Managing Digital Collections:
A Collaborative Initiative
on the South African Framework

Acknowledgements

Contents

Introduction

Developments in information and communication technologies (ICTs) have presented opportunities for the rapid production of data, digital content, digital collections, institutional and subject repositories, digital libraries and archives. Some of the new areas of decision-making faced by directors, curators and managers are setting up a networked infrastructure, purchase of hardware and software, copyright concerns, workflow and quality assurance.

Developing countries like South Africa are following the digitisation trend set by developed countries. Many organisations need to go ahead with new digital projects despite financial constraints and diminishing institutional budgets. Creating, managing, sharing and preserving digital content in a responsible and sustainable manner creates new challenges, demands and requirements. It requires *inter alia*

- expertise
- human and financial resources
- technology infrastructure
- adoption of standards
- creation of guidelines
- implementation of policies

It can be a daunting task to make long-term decisions, create budgets and financial plans without an overall understanding of the digital landscape, the requirements and implications of creating digital resources as well as the long term responsibilities that are inevitable.

1.1 Background and overview

In February 2009 South African practitioners and leaders in the field of digitisation met to find ways of accelerating the development of digital collections as well as increasing the scope and extent of digitised South African resources available on the Web. At this meeting, it was agreed that the collaborative production of a reference document for managers level would provide a valuable framework for organisations considering the digitisation of their resources. The document would provide an overview of requirements for building good digital collections, sound digital collection management practices and guidelines for data sharing and long term preservation and access.

Managing Digital Collections: A Collaborative Initiative on the South African Framework is the publication that has resulted after many months of research, writing, reviewing, editing and finally web publishing.

Each chapter is devoted to a different area of the digitisation process and is authored by an authority in that area of specialisation. A chapter on community practice examines the emerging community developing standards and guidelines in order to find consensus about what constitutes good practice. It reviews at documents such as reports that have shaped and continue to shape principles of good practice. Before proceeding with any digitisation, the intellectual property issues and most specifically copyright issues need to be understood within the legal framework.

There is a chapter on the current South African legislation and its relation to digital resources.

Digital collection development is the broad term for the management of digital collections which may not necessarily be housed in any one location but may consist of several collections on a theme, across several organisations. A chapter on selection criteria and development criteria is complemented by a subsequent chapter on digital objects.

Metadata plays an essential role in the access, storage, discovery, preservation and exchange of digital resources. It provides the mechanism for efficient and effective management of all digital resources. A chapter on metadata examines international standards for the creation of quality metadata.

A chapter on digital infrastructure focuses on the hardware, software and middleware required to host digital collections as well as the digital content management system to ingest, preserve and deliver content to and from the Web.

Components required for an efficient digital preservation program are discussed in a chapter where the resources required to sustain preservation programs are discussed. The final chapter of the publication addresses the planning, overall management, quality assurance and evaluation of a digital collections project.

The Framework of Guidance for Building Good Digital Collections, accessible on the NISO website (http://www.niso.org), is a good guide document and is referred to and is used in several chapters of the South African Framework. This reiterates the importance of adopting internationally recognised principles, where possible, for local requirements.

1.2 Objective

The objective of this Framework is to provide high-level principles for planning and managing the full digital collection life cycle. It aims to
- provide an overview of some of the major components and activities involved in creating good digital collections
- provide a sense of the landscape of digital collections management
- identify existing resources that support the development of sound local practices
- encourage community participation in the ongoing development of best practices for digital collection building
- contribute to the benefits of sound data management practices, as well as the goals of data sharing and long term access
- introduce data management and curation issues
- assist cultural heritage organisations to create and manage complex digital collections
- assist funding organisations who wish to encourage and support the development of good digital collections
- advocate the use of internationally-created appropriate open community standards to ensure quality and to increase global interoperability for better exchange and re-use of data and digital content.

The framework provides principles and guidelines to assist planners, policy developers and managers of digital collections gain a sense of the digitisation landscape. It can be either a reference (to be consulted as and when required) or an overview (to be read from beginning to end). It is not a step-by-step manual of digitisation processes – that will be the function of training initiatives for personnel actually doing the digitisation – but it can support and supplement training by being prescribed reading for digitisation students and staff. This will assist them by providing an overall understanding of the broader perspectives and the multiple processes involved in building good digital collections.

Authoritative references to assist in conducting further research are provided at the end of each chapter. Other useful resources (including websites), commonly used acronyms and a glossary of terms are provided at the end of the Framework.

The framework is not intended to be prescriptive but rather to provide information to support individual organisational decision-making best suited to their own collections and specific needs. Specific technical specifications change rapidly and are therefore not included.

1.3 Intended audience

The document is directed at organisations that are or will be developing digital collections and specifically managers at those organisations who will be responsible for managing such collections. It is intended as a reference source for decision-makers such as digital project managers, repository managers, curators, digital information managers, senior librarians and information technology managers.

The framework is broad enough to accommodate different organisations, varying skills and budgets while still being specific enough to provide sufficient detail to support decision making in a variety of circumstances.

It is hoped that this publication will be adopted as a definitive reference tool and will contribute to expanding the nature and extent of digitisation efforts in South Africa. Organisations are urged to document their practices and decisions in guidelines for use in their own environment, thereby ensuring consistency, quality assurance and ongoing good practice.

Authors' names and affiliations (in alphabetical order)

Ria Groenewald
Department of Library Services
University of Pretoria

Wouter Klapwijk
Library and Information Service
Stellenbosch University

Patricia Liebetrau
Digital Innovation South Africa (DISA)
University of KwaZulu-Natal, Durban

Nancy Y. McGovern
Director, Digital Preservation Management Workshops
Hosted by the Inter-university Consortium for Political and Social Research (ICPSR)
University of Michigan

Denise Rosemary Nicholson
Copyright Services Office
The Library
University of the Witwatersrand, Johannesburg

Felix N Ubogu
University Librarian
The Library
University of the Witwatersrand, Johannesburg

Chapter 1

Community practice

1. Introduction

The digital preservation community includes the international combination of individuals and organisations that are committed to providing access to digital content across generations of technology. Digital content may be created in digital form – *born digital* – or transformed into digital form – *digitised*. An organisation that takes on responsibility for managing digital collections over time – e.g. an organisation that is creating digital content through digitisation – will benefit from an awareness of the community landscape in which digital collections are managed. The digital community is an emerging community that is developing standards and consensus regarding good practice. Digital content has been preserved since the 1960s by archival programs at data archives and national archives, since the 1990s by libraries with growing collections of digital images and growing amounts of born-digital content, and increasingly since the early 2000s by museums, which are increasingly responsible for preserving digital art. Since the mid-1990s, these distinct domains have been coming together with the common objective of preserving digital content for use by current and future users. This chapter provides a brief overview of the emergence of the digital community and highlights core community standards and practice.

Within an emerging community, the language to be used by members needs to be developed as standards and practice are formalised.[1] The term *digital preservation* encompasses all of the activities that are undertaken by digital collection managers to ensure that the digital content is maintained over time in usable formats and can be made available in meaningful ways to current and future users. *Digital stewardship* includes the set of actions taken by caretakers or curators to extend the longevity and usefulness of digital content. Lifecycle management refers to the sequence and iteration of actions that begins with the creation of digital content and continues through the long-term management, use and re-use of content to the ultimate disposal of the portion of content that is no longer needed. The term *data curation* refers to activities curators engage in that add value to digital content to make it meaningful or useful to the users that create and/or rely upon it. In practice, data curation may refer to specific types of research data or to digital content more generally, depending on the context. The term *digital curation* was introduced and promoted by the Digital Curation Centre in the United Kingdom encompasses data curation and digital preservation.

[1] Some core concepts that are referred to in this chapter are being refined by the community. This chapter reflects broad-based community understanding of relevant terms at the time of its publication.

Since 1995, community efforts have defined standards for managing digital content. This chapter reviews the six documents that have been developed and accepted by the community and comments on the potential benefits of each for digital content managers. These documents cumulatively reflect the current set of standards for the long-term management of digital collections. Documents that were developed for specific communities, e.g. the records management standard [ISO 15489-1], or that are not generally relevant to managing digital collections, are not included

2. Preserving digital information, 1996

The emergence of the digital community may be traced to the publication in 1996 of the report *Preserving Digital Information: Report of the Task Force on Archiving of Digital Information* report. It is an example of a community effort that combined expertise from the domains of library, archives and museums, a collaborative approach to addressing the challenge of digital preservation that had not occurred prior to the development of this report. The Commission on Preservation and Access (now the Council on Library and Information Resources) and the Research Library Group in North America convened a task force to address the challenge of digital preservation across the communities that maintain digital collections. *Preserving Digital Information* is a very readable document that identifies four challenges faced by organisations in managing digital collections over time – i.e. avoiding technological obsolescence; migrating digital content over generations of technology; addressing relevant legal and organisational issues; and providing community-wide infrastructure for managing content. It also provided recommendations for the community to address these challenges. More than a decade later, some of the challenges have been addressed and all remain relevant.

The report defined the following five integrity features that managers of digital collections must address:

- *Content* requires that the essence of the content should be preserved. For digital image collections, this means ensuring that the images completely and correctly reflect the content that was digitised.
- *Fixity* requires that any changes to the object should be detected, recorded and reversed or corrected, if appropriate. For digital images, the collection manager might ensure that a checksum is generated for each image and used to detect and avoid errors.
- *Reference* requires that each digital object should be able to be uniquely identified and cited as distinct from other digital objects for the lifetime of the objects. For digital images, a collection

manager might adopt an existing identification scheme, such as handles.[2]

- *Provenance* requires that the digital content should be traceable to its origin or inclusion in a managed digital collection. For digital images, the provenance might be noted at the collection level (e.g., what content was digitised, by whom, when, and how), if the information applies to all digital objects in the collection.
- *Context* requires that linkages with other objects, dependencies on specific technologies, dissemination restrictions, and the social setting of the digital object should be preserved as appropriate. For digital images, it is essential to identify and enable linkages to other digital objects and to avoid technology dependencies by using common and readily-available formats (e.g., TIFF, JPEG, PDF).

These integrity features are essential for the long-term management of digital collections and are fairly easy to address for digital image collections, as suggested by the examples provided.

3. Open archival information systems (OAIS), 2002

The Open Archival Information System (OAIS) Reference Model [ISO 14721: 2003] has had an enormous impact on the digital community. The Consultative Committee for Space Data Systems (CCSDS) began this initiative in 1995 to meet its own space data management requirements and as a service to the community. The initiative included an international group of experts engaged in digital content management across domains, such as libraries, archives, and museums. OAIS was informed by the *Preserving Digital Information report*.[3] For example, the OAIS Reference Model incorporates in its definition of the contents of an Archival Information Package (AIP) the five integrity features from that 1996 report. The purpose of the development of the OAIS standard was to raise awareness about archival concepts, to enable broad and effective participation in the preservation process, to enable organisations to describe and compare what they do (operations) and how they do it (architecture), to enable the specification of preservation strategies and approaches, to enable the definition of models for every kind of digital content, and to encourage the development of OAIS-related standards. The OAIS Reference Model was completed in 2002 and approved by the International Standards Organization (ISO) in 2003. The five-year revision of the OAIS standard was released for public comment in May 2009 with few substantive changes. For digital collections management, it is worth noting that access rights information was added to the components of the AIP in the 2009 revision.

OAIS was developed to be applicable in any organisational context in which digital content is managed for the long-term. The promulgation of the OAIS standard has provided a common language for the digital community. The majority of organisations that manage digital collections have indicated an intention to design and implement their digital repositories in accordance with OAIS. Two of the most commonly-used open source repository systems, Fedora and DSpace, are working towards integrating OAIS concepts and principles, making it easier for digital collection managers to comply with the community standard.[4] There are a number of OAIS-related developments that are useful for digital collection management, including persistent identifier schemas, preservation metadata (e.g. Preservation Metadata: Implementation Strategies or PREMIS), protocols for the audit of digital archives, and the Producer-Archive Interface Method Abstract Interface (PAIMAS), that was completed in 2004 and approved by ISO in 2006 [ISO 20652: 2006].

4. Trusted digital repositories (TDR), 2002

The *Attributes of a Trusted Digital Repository: Roles and Responsibilities* report examined the requirements an organisation would have to meet to conform to OAIS. Research Library Group (RLG) and the Online Computer Library Center (OCLC), which were merged in 2007, convened a group of international experts to develop the TDR document. The document defines seven attributes of a trusted digital repository, an organisation that is committed to preserving designated digital content over time:

- *Administrative responsibility:* an explicit commitment by an organisation that takes responsibility for preserving digital content;
- *Organisational viability:* the wherewithal of an organisation to do digital preservation, e.g., the legal status, policies, procedures, and skills (see Chapter 7);
- *Financial sustainability:* adequate funding that is designated by an organisation to engage in digital preservation;
- *Technological and procedural suitability:* appropriate approaches have been identified to preserve and manage the digital content over time;
- *System security:* adequate control of access to and potential vulnerabilities of digital collections;
- *Procedural accountability:* documented approval and implementation of decisions, policies, procedures, and practices;
- *OAIS compliance:* a commitment to develop approaches in accordance with the ISO standard in designing and implementing lifecycle management programs.[5]

---

2 The Handle System, http://www.handle.net/.
3 "ISO Archiving Standards – Overview," http://nssdc.gsfc.nasa.gov/nost/isoas/overview.html.
4 Fedora Commons, http://www.fedora-commons.org/, and DSpace Federation, http://www.dspace.org/.
5 The discussion of digital preservation in Chapter 7 uses these attributes as a framework for characterising an organisation's digital preservation program.

Prior to the release of TDR in 2002, the cost of managing digital content (financial sustainability) had not been referenced so explicitly within the community. Technology was typically the first topic of digital preservation discussions, and the need to demonstrate through organisational evidence the adequacy of an organisation's operations had not been required. In addition to the attributes, the TDR document addresses the nature of trust for organisations that manage digital content over time. TDR has had a significant community impact.

5. Trustworthy Repositories Audit and Certification (TRAC), 2007

With the release of the TDR document and the approval of OAIS as an international standard, the community had two foundation documents for digital preservation that defined the organisational context for managing digital collections over time (i.e. TDR) and described the technological context for a system to support lifecycle management of digital collections (i.e. OAIS).[6] These two documents formed the basis for the development of the *Trustworthy Repositories Audit & Certification (TRAC): Criteria and Checklist* by the Research Libraries Group and National Archives and Records Administration (RLG/NARA) Task Force on Digital Repository Certification. Although the effort was charged with developing the means for community-based certification, the focus shifted to the definition of criteria for self-assessment of a digital archive and the possible audit by external auditors and peers with the potential of an eventual international certification programme. TRAC is a fundamental component of a series of international initiatives on audit and certification that includes developments in the United Kingdom and Europe (DRAMBORA), a programme in Germany (nestor), and the development of an ISO standard for audit and certification that is based on TRAC.[7] The working group for the standard used the TRAC requirements as a starting point beginning in 2007 and submitted a revision of the requirements in spring 2009. The revised set of TRAC requirements will be included in the standard when it is approved. The revised version reorganises the TRAC requirements, but the core requirements remain for three sections of activities:
- *Organisational Infrastructure*, including governance and finance;
- *Digital Object Management*, including all stages of life cycle management;
- *Technologies, Technical Infrastructure, and Security*.

Many organisations have been using the requirements for self-assessment and audit of their programmes. The self-assessment process produces a development and improvement plan that is invaluable for effective digital collections management.

6. A framework of guidance for building good digital collections, 2004-2007

The National Information Standards Organization (NISO) released the third version of *A Framework of Guidance for Building Good Digital Collections* in 2007. This community document defines a useful set of practical principles on four aspects of digital collection management:

- *Collections:* aggregations of digital objects;
- *Objects:* individual packages of digital content to be managed, preserved, and used;
- *Metadata:* information about digital content that enables and enhances use;
- *Initiatives:* activities within digital programmes to encourage quality content creation and use.

Any organisation with responsibility for digital collections should be aware of these principles that encourage sound, community-based practice. The guidance was initially developed through a grant funded by the Institute of Museum and Library (IMLS) then adopted by NISO, making the principles more widely available within the digital community.

7. Data seal of approval

The *Data Seal of Approval* is a concise set of higher-level principles for producers of digital content, repositories that manage digital content, and consumers of digital content. The initial set of principles was developed and released by the national data archives in the Netherlands, DANS, in 2008. The Data Seal of Approval has since been adopted for international use by data archives and is being extended to apply to digital content of any kind. The principles reflect the fundamentals of both TDR and OAIS and represent a manageable approach to digital content management.

8. Good practice for digital collections

There are numerous standards that pertain to digital collections management, but the community documents discussed in this chapter represent a core set for digital collection managers to address.

Figure 1.1. Value of community documents for digital collection management.

| Community Document	Value for Digital Collection Management
Preserving Digital Information	Provides a framework for issues and challenges
OAIS	Describes technology infrastructure for managing content
TDR	Characterises organisational issues for managing content
TRAC	Provides requirements for progress and compliance
NISO Good Digital Collections	Recommends practice for creating and managing content
Data Seal of Approval	Defines high-level principles and an emblem for compliance

---

One of the challenges for digital collection managers is to monitor and respond to ongoing developments in the community. As Figure 2 illustrates, community practice may emerge from top-down initiatives, e.g. standards development, or from bottom-up applications, e.g. local practice in one or more organisations. Good practice within the community evolves over time as new requirements are defined, new types of digital content are developed, and organisations take responsibility for solving problems through local application of community standards and other developments.

Figure 1.2. The evolution of community standards and practice.
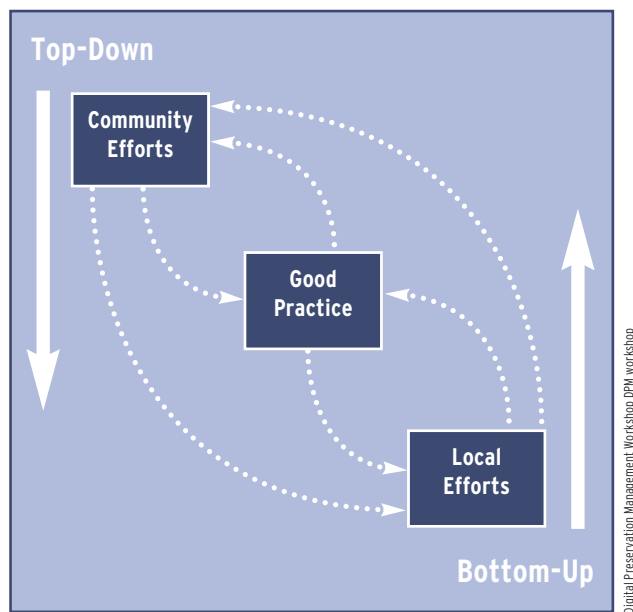
Over the past five years, international conferences have been established that provide updates through their online proceedings. The Society for Image Science and Technology (IS&T) Archiving Conference, which emerged from the digital imaging community, has been held since 2004.[8] The International Conference on the Preservation of Digital Objects (iPres), which addresses the preservation of any kind of digital content, has been held annually since it began in 2004.[9] In addition, the National Library of Australia, now in collaboration with the Digital Preservation Coalition in the UK, has hosted the Preserving Access to Digital Information (PADI) Web site, a central source of information on all aspects of digital preservation. These are good examples of a growing number of resources for tracking developments within the digital community. Digital collection managers have a foundation of good practice on which to build.

9. Conclusion

The need to standardise good practices for digital preservation has resulted in the designing of guidelines for various aspects of creating, managing and tracking necessary procedures. The international cooperation between systems and people needed to produce and support a task of this magnitude has led to the establishment of a community that interacts with members for the benefit of all.

References

Consultative Committee for Space Data Systems (CCSDS). Recommendation for Space Data System Standards: Reference Model for an Open Archival Information System (OAIS). Blue Book CCSDS 650. 0-B-1, no. 1 (2002).
http://public.ccsds.org/publications/archive/650x0b1.pdf.

Data Seal of Approval, 2008. http://www.datasealofapproval.org/

International Standards Organization. ISO 14721:2003: *The Open Archival Information System (OAIS) Reference Model*, Geneva, Switzerland: International Standards Organization, 2003.

National Information Standards Organization (NISO). *A Framework of Guidance for Building Good Digital Collections*, 2007.
http://www.niso.org/publications/rp/framework3.pdf

Research Library Group (RLG) and Online Computer Library Center (OCLC). *Trusted Digital Repositories: Attributes and Responsibilities*. Mountain View, CA: Research Library Group (RLG), 2002.
http://www.oclc.org/programs/ourwork/past/trustedrep/repositories.pdf

Research Library Group (RLG)-National Archives and Records Administration (NARA) Task Force on Digital Repository Certification. *Trustworthy Repositories Audit & Certification: Criteria and Checklist* Ver. 1.0. Chicago, IL: Center for Research Libraries (CRL), 2007.
http://www.crl.edu/content.asp?l1=13&l2=58&l3=162&l4=91

Waters, Don, and John Garrett, *Preserving Digital Information. Report of the Task Force on Archiving of Digital Information*. Washington, DC: The Commission on Preservation and Access and Research Libraries Group, 1996. http://www.clir.org/pubs/reports/pub63watersgarrett.pdf

---

[8] There is a Web site for all IS&T conferences that includes the archiving conferences. Society for Imaging Science and Technology (IS&T), 'IS&T Meetings Calendar', http://www.imaging.org/conferences/recentmeetings.cfm (accessed 2 April 2008).
[9] International Conference on Digital Preservation (iPres), http://rdd.sub.uni-goettingen.de/conferences/ipres/ipres-en.html (accessed 2 April 2008).

Chapter 2

Copyright and related matters

1. Introduction

Copyright is a very important aspect of the digitisation process, as it encompasses the legal considerations that have to be made regarding the creation and maintenance of digitised collections. Briefly, a copyright is a right granted by law to an author, designer, or artist to prohibit others from copying or exploiting his or her works in various ways without permission. This chapter will provide an introductory guide to assist librarians to interpret and apply the copyright law when they manage digital collections. Directly or indirectly, copyright shapes the content of digital collections. For more detailed information on this topic, readers should consult the South African Copyright Act No. 98 of 1978 (as amended) and other references provided at the end of this chapter. Should any legal advice be required, an institutional legal advisor or intellectual property lawyer should be consulted.

1.1 Definition and governance

Copyright is a category of intellectual property which represents the property of the mind or intellect. It is a statutory monopoly or a 'bundle' of exclusive rights conferred by the law on authors and creators to protect their original works. It is governed by the South African Copyright Act No. 98 of 1978 (as amended) (hereinafter called the "Copyright Act")[10]. South Africa is a signatory to various international intellectual property agreements, for example:
- The Berne Convention[11];
- World Trade Organisation's (WTO) Trade-Related Aspects of Intellectual Property Rights Agreement ("TRIPS")[12];
- The World Intellectual Property Organisation (WIPO) Copyright Treaty ("WCT") [13] (not yet ratified);
- The World Intellectual Property Organisation (WIPO) Performances and Phonograms Treaty ("WPPT")[14] (not yet ratified).

South Africa is obliged to adopt the minimum requirements or standards in these agreements in its national law, and to afford the same protection to foreign authors and creators in signatory countries, as it does to South African authors and creators.

1.2 Category of works protected

Under the South African Copyright law, the following works are protected:
- Literary works;
- Musical works;
- Artistic works;
- Cinematographic films;
- Sound recordings;
- Broadcasts;
- Programme-carrying signals;
- Published editions;
- Computer programs.

1.3 Copyright ownership

Subject to Section 21 of the Copyright Act, there are various categories of ownership. These are as follows:
- Copyright ownership in literary and musical works vests in the author or creator, or in the case of joint authorship, in the co-authors of the work.
- Artists, sculptors and cartoonists hold copyright of their artistic works.
- In the case of photographs, copyright belongs to the person responsible for the composition of the photograph.
- The person who arranges for the making of sound recordings or films holds copyright.
- The publisher owns copyright of published editions.
- In the case of a computer-generated literary or artistic work, copyright is owned by the person who undertakes the arrangements necessary for the creation of the work.
- Copyright in a broadcast belongs to the person who first broadcasts it.
- Concerning programme-signals, copyright is owned by the first person to emit the signal to a satellite.
- Copyright in Government publications, online services and Government websites belongs to the State. However, official texts of a legislative, administrative or legal nature or individual political speeches or in speeches delivered in the course of legal proceedings, or translations thereof, are in the public domain and are excluded.

---

[10] http://www.gpa.co.za/pdf/legislation/Copyright%20Act.pdf
[11] http://www.wipo.int/treaties/en/ip/berne
[12] http://www.wto.org/english/tratop_e/trips_e/trips_e.htm
[13] http://www.wipo.int/treaties/en/ip/wct
[14] http://www.wipo.int/treaties/en/ip/wppt

1.4 Transfer of copyright

Copyright can be transferred to other parties as movable property by assignment, testamentary disposition or operation of law. The following points are adapted from the South African Copyright Act No. 98 of 1978 (as amended):

- When a literary or artistic work is made by an author and published or reproduced in a newspaper, magazine or similar periodical in the course and scope of being employed under a contract of service or apprenticeship by the proprietor of the publication, the proprietor (and not the author) is the copyright owner.
- When a person commissions the taking of a photograph, the painting or drawing of a portrait, the making of a gravure, the making of a cinematograph film or the making of a sound recording, and pays or agrees to pay for it in money or money's worth, and the commission is fulfilled, the person who has commissioned the work is the copyright owner, subject to the provisions of the sub-first paragraph above, by virtue of Section 3 and 4 of the Copyright Act.
- Where the first sub-paragraph or the second paragraph above do not apply and a work is made in the course of the author's employment by another person under a contract of service or apprenticeship, that other person (the employer) shall be the owner of any copyright subsisting in the work by virtue of Section 3 or 4 of the Act

In many South African tertiary institutions, the institution holds copyright of research reports, theses and dissertations by virtue of institutional policies and written assignments to the institution by staff and postgraduate students. If the work is done in the course and scope of a staff member's employment, the tertiary institution, and not the staff member, then holds copyright in those works.

1.5 Exclusive rights of authors and creators

There are specific exclusive rights depending on the category of copyright work. Exclusive rights given to authors and creators include the reproduction of their works in any manner or form, for instance publishing their works; performing their works in public; broadcasting them or causing them to be transmitted in a diffusion service; making an adaptation or including the works in other works, e.g. including an artistic work in a cinematographic film or a television broadcast.

Authors also have moral rights in terms of Section 20 of the Copyright Act. Moral rights give authors the right to claim authorship of the work and to object to any distortion, mutilation or other modification of the work where such action is or would be prejudicial to the honour or reputation of the author(s). These rights are not transferable.

1.6 Limitations and Exceptions to Exclusive Rights of Authors and Creators

In terms of 'Section 12 of the Copyright Act, a 'fair' portion of a work, may be copied in terms of '*Fair Dealing*', without permission for the following purposes:

- Research or private study;
- Personal or private use;
- Criticism or review;
- Reporting current events (e.g. in a newspaper or broadcast).

Fair dealing also applies to other works such as artistic works, films, sound recordings, broadcasts, published editions and computer programs.

Section 12 (2-4) of the Copyright Act also allows copying, *without permission*, for the following purposes:

- Judicial proceedings; or a report of such proceedings;
- Quotation;
- By way of illustration for teaching purposes (e.g. placing an extract of a work on an overhead projector, or in PowerPoint presentation, to highlight aspects of a lesson in progress).

Section 13 Regulations of the Act allow limited copying, *without permission*, for teaching purposes (e.g., single handouts for students in a classroom situation, single copies for lecturers for research, teaching or preparation for teaching in a class). These provisions do not extend to distance or open learning, literacy training programmes, adult basic education training (ABET), non-commercial staff training programmes, or digitisation of works.

Section 13 Regulations also have limited reproduction exceptions for libraries and archives. Although the legislation is not media-specific, these exceptions are not practical in the digital environment.

1.7 Term of Copyright Protection

South Africa adopted into its Copyright law the minimum period of copyright protection required in international copyright agreements. The following terms of protection apply:

a) Literary and musical works or artistic works, other than photographs are extended for the lifetime of the author plus 50 years from the end of the year in which the author dies. This is provided that if, before the death of the author, none of the following acts had been done in respect of the works or adaptations:
- The publication thereof;
- The performance thereof in public;
- The offer for sale to the public of records thereof;
- The broadcasting thereof.

The term of copyright shall continue to subsist for the period of 50 years from the end of the year in which the first of the said acts is done.

b) Cinematographic films, photographs and computer programs extend for 50 years from the end of the year in which the work is made available to the public with the consent of the owner of the copyright, or, failing such an event within 50 years from the making of the work, 50 years from the end of the year in which the work is made.

c) Sound recordings extend for 50 years from the end of the year in which the recording is first published.

d) Broadcasts extend for 50 years from the end of the year in which the broadcast first takes place.

e) Programme-carrying signals extend for 50 years from the end of the year in which the signals are emitted to a satellite.

f) Published editions extend for 50 years from the end of the year in which the edition is first published.

g) Literary, musical or artistic works (other than photographs) made by or under the control of the State extend for 50 years from the end of the year in which the work is first published.

h) Copyright protection for other categories of works (i.e. other than those in (g) above) made by or under the control of the State is subject to the same term of copyright provided for in (a)-(f) above, depending on the type of work.

In terms of the Copyright Act, works are released into the public domain when the copyright term expires. Rights holders can, however, release their works into the public domain earlier, if they so wish.

As mentioned above, there is no copyright in Government official texts of a legislative, administrative or legal nature, or in official translations of such texts. Similarly, there is no copyright in speeches of a political nature, or in speeches delivered in the course of legal proceedings. The author of political speeches, however, has the exclusive right to make a collective work of his/her political speeches.

There is no copyright in abstract ideas (only in the expression of those ideas), facts, or news of the day that are mere items of press information.

1.8 Infringement

Copyright is said to have been infringed when a 'copy' or 'reproduction of the work' has been made in the sense that it is substantially similar to a copyright work, and it must have been copied from the copyright work as opposed to being the result of coincidental independent production or having been taken from the same sources as the copyright work.[15]

2. Copyright in the digital environment

Digitisation is a process of converting printed works into binary or numeric machine-readable code. Unlike a photocopy, a digitised version of a work involves more than reproduction. It involves the conversion to another format, often involving modification, adaptation or crop-

ping, even translation, where necessary. It enables a work to be searched, browsed, amended, and enhanced by any number of users at the same time. It is a means of making information electronically accessible to a wide audience, in other words, it is a form of publishing.[16] Anyone who places information on the open Web is essentially an online publisher. Altering someone else's work, like leaving out sections of a work too fragile or too complicated to be digitised, cropping information or photographs, or removing graphics or advertisements from a work, may constitute an infringement of the author's or the creator's moral rights.

The intimate connection between access and copying has considerable significance in the context of copyright protection. One of the essential elements of copyright is the right to control reproduction. In the digital world, access is not possible without copying. By merely browsing an item on a website, transient or temporary copies are always created in the process.[17]

2.1 Using public domain material in digital collections

If public domain material is included in a digital project without any changes (i.e. it is merely a digital copy of the same information), then no copyright costs should be charged for the use of that material. This is because the copyright has been exhausted and the rights owners would already have benefited economically during the copyright term. However, users may be charged for access and related administrative costs.

If the work in the public domain is enhanced to create a derivative work, or a new digital work which meets the requirements for protection under the Copyright Act, then copyright would subsist in that new product. This would apply even if the new product includes infringed works or is an infringement in itself, e.g. a translated work which has been done without permission of the rights owner.

2.2 Copyright provisions for preservation, digitisation and digital curation

There are limited exceptions for libraries with regard to *preservation* in the *analogue* environment, but not in the *digital* environment. These include:

- Clause 3(d) of Section 13 Regulations reads: 'the rights of reproduction and distribution shall apply to a copy of an unpublished work duplicated in facsimile form solely for purposes of preservation and security or for deposit, for research use, in another library or archive depot, provided that the copy reproduced is to be placed in the collection of the library or archive depot";
- Clause 3(e) reads: "the right or reproduction shall apply to a copy of a published work duplicated in facsimile form solely for the purpose of replacement of a copy that is deteriorating or that has been damaged, lost, or stolen: Provided that the library or archive depot has, after a reasonable effort, determined that an unused replacement cannot be obtained at a fair price."

There are no provisions in the Copyright Act for reproduction or preservation of whole collections, or making works public, i.e. multiple

---

[15] http://www.america.gov/st/business-english/2008/January/20080109153644AKllennoCcM0.3526117.html

[16] Buys. R.

[17] Longe, O.B.

copying or publishing works to a wide audience, as would occur if they were made accessible in a digital form. Its underlying requirement that a published work can only be reproduced if an unused replacement cannot be obtained at a fair price, has cost and other implications for libraries, particularly when engaging in large scale preservation or digitisation projects. Libraries are obliged to purchase unused replacements or to seek copyright clearance for each work. Libraries and archives therefore need to budget for the repurchase of material or for large-scale copyright clearance costs.

2.3 Copyright provisions for interlibrary loans/document delivery services

Sub-clauses 3(f)-(h) of Section 13 (Regulations) permit a library or archive to reproduce and distribute single copied items made from its library, or obtained from another library, to users at their request, provided that the items become the property of the users and the work is only for private study or personal or private use of the persons using the works. This generally applies to *analogue* works.

In *digital* collections, provision of document delivery services will be determined by specific licence conditions related to digital works, or the terms of use permitted by rights holders when granting permissions.

2.4 Copyright provisions for conversions into alternative formats

There are no provisions in the copyright law for conversions into alternative formats, for instance Braille for blind persons, or into more visual formats for deaf persons. Libraries need to clear copyright before making any conversions of copyright works for persons with sensory-disabilities.

When planning a digital project, a library must include issues affecting persons with sensory-disabilities in its policy and procedures. Persons with sensory disabilities include those who are blind, visually or perceptually impaired, blind and deaf, deaf or partially deaf, dyslexic, or those who suffer from physical, learning or other disabilities which might affect their access to digital works.

All material should be accessible, without restriction or prohibition by digital rights managements or systems or technological protection measures. Libraries should be given the 'keys' to decode or decrypt such protection measures to enable legitimate access to such works by **all** users, including those with sensory disabilities, particularly in accordance with Sections 12 and 13 of the Copyright Act.

2.5 Digital rights management systems or technological protection measures

The Copyright Act does not address digital rights management systems (DRMs) or technological protection measures (TPMs). However, Clause 86 of the Electronic Communications and Transactions Act No. 25 of 2002 provides for anti-circumvention measures and the prohibition of devices to circumvent such measures. However, it does not have any exceptions for 'fair dealing', legitimate library functions or access by persons with sensory-disabilities. This clause

prohibits users from bypassing or circumventing copyright technological protection measures or digital rights management systems, even for legitimate purposes, and effectively 'locks up' information indefinitely.

Digital rights management (DRM) is a generic term that refers to access control technologies that can be used by hardware manufacturers, publishers, copyright holders and individuals to try to impose limitations on the use of digital content and devices. Technical protection measures (TPMs) are technological tools used to restrict the use and/or access to a work.[18]

3. Licences in the digital environment

In the digital environment, contract law tends to override copyright exceptions. Licences for electronic works specify exact terms of use and re-use, even if they are stricter than provisions permitted in the Copyright Act. When using material from electronic databases in digital projects, the terms of the licence need to be adhered to. If the licence does not permit digitisation or inclusion in a digital collection, then prior permission has to be sought and fees paid to the rights holders if required.

Electronic databases do not have standard licence conditions. Each database provider or rights holder has its own specific e-licences setting out rules and conditions relating to the use or re-use of its digital content.

Many authors and creators now use Open Source licences to make their works more accessible in the digital environment. These licences are more flexible and user-friendly; for example the *GNU Free Document License*[19] and *Creative Commons*[20] In South Africa, these licences are based on the South African Copyright Act No. 98 of 1978 (as amended) and are irrevocable, for example a work that is subject to a non-commercial licence cannot be used for commercial purposes. If works under Creative Commons licences are used in digital collections, the conditions of their respective licences must be followed and perpetuated by all users.

3.1 Creative Commons Licences

Creative Commons have a number of different types of licences. These are:
- **Attribution**
  This licence allows users to distribute, remix, tweak and build upon someone else's work, even commercially, as long as credit is given to the author of the original creation. This is the most accommodating of licences offered, in terms of what users can do with a work licensed under Attribution.
- **Attribution Share Alike**
  This licence allows users to remix, tweak and build upon someone else's work, even for commercial reasons, as long as credit is given to the author and they licence their new creations under the identical terms. This licence is often compared with open source software licences. All new works based on the original licensed work will carry the same licence, so any derivatives will also allow com-

[18] http://en.wikipedia.org
[19] http://www.gnu.org/copyleft/fdl.html
[20] www.creativecommons.org

mercial use.

- **Attribution No Derivatives**
  This licence allows for redistribution (commercial and non-commercial), as long as it is passed along unchanged and as a whole, with credit given to the original author.
- **Attribution Non-Commercial**
  This licence allows users to remix, tweak and build upon the original licensed work non-commercially, and although their new works must also acknowledge the original author and be non-commercial, they do not have to license their derivative works on the same terms.
- **Attribution Non-Commercial Share Alike**
  This licence allows users to remix, tweak and build upon the original licensed work non-commercially, as long as they give credit to the author and license their new creations under the identical terms. Users can download and redistribute the licensed work just as with the *Attribution Non-Commercial No Derivatives licence* (see below), but they can also translate, make remixes and produce new stories based on the licensed work. All new work based on theirs will carry the same licence, so any derivatives will also be non-commercial in nature.
- **Attribution Non-Commercial No Derivatives**
  This licence is the most restrictive of the six main Creative Commons licences that allow redistribution. This licence is often called the "free advertising" licence because it allows users to download the licensed works and share them with others as long as they mention the author and link back to the author, but they may not change the works in any way or use them commercially.
- **Creative Commons Zero (CC0)**
  Copyright and other laws throughout the world automatically extend copyright protection to works of authorship and databases, whether authors or creators want those rights or not. CC0 now gives people who want to give those rights away to do so to the fullest extent allowed by law. Once the creator or a subsequent owner of a work applies CC0 to a work, it is no longer his or hers in any meaningful legal sense. Anyone can then use the work in any way and for any purpose, including commercial purposes, subject to rights others may have in the work, or how the work is used.

  CC0 provides a "*no rights reserved*" option for authors and creators wanting to release their works into the public domain.

  Libraries need to be aware of the licence terms and conditions of all licences before including them in a digital collection. Licence conditions need to be adhered to and perpetuated when material is converted or migrated to new formats.

4. Copyright permissions

Libraries need to prioritise which works or collections require digitisation, or which works require migration for digital curation purposes. The following are examples of what should be taken into consideration when decisions are made: public domain material should be considered first; the age and value of works; whether items or works are fragile, damaged or of cultural importance; the topical relevance or importance of the works; whether technological upgrades are necessary and what technology is available for migration.

4.1 Priorities and procedures with regard to copyright clearances

Libraries need to decide on and prioritise which works require copyright clearance. They should do the following:

- Establish whether the works are in the public domain. If not, they need to establish who owns the copyright, e.g. individuals, institutions, organisations, shared or joint owners (known and anonymous), research organisations or funding agencies, and so on.
- Approach the relevant copyright holders. The Dramatic, Artistic and Literary Rights Organisation (DALRO) have a mandate to clear only reprographic reproductions and transient electronic copies. Permissions for works to be digitised or to convert, adapt, translate or migrate born-digital works need to be obtained directly from the rights holders.
- Establish whether the work has more than one copyright holder, e.g. a film, video or DVD can incorporate a number of different copyright works. Permission would be needed from all relevant copyright holders.
- Establish whether all parts of multimedia can be made accessible or whether there are embargoes on some.
- Establish whether there are any digital rights management systems with technological protection measures embedded in the works to be digitised, or in the born-digital works. The library would need to obtain the 'keys' or decryption codes from the rights holders to 'unlock' the content in order to enable access to these works and/or to engage in preservation or digital curation activities.
- Establish whether the work is 'orphaned'. When the rights holders are difficult to identify, contact or find, or are untraceable, their works fall into the category of '*orphan works*'. Decisions on how to deal with orphan works need to be made by the library and project team.
- There are three lines of action to take. These are:
  i. Engage in a reasonable search to find the copyright holder to obtain permission;
  ii. Abandon use of the material if permission cannot be obtained;
  iii. Proceed with digitisation of the material (depending on its importance and value) but provide a disclaimer and invitation to rights holders to negotiate a reasonable copyright fee.

    It is advisable to make every effort to obtain permission from rights holders and to keep records of these attempts, before deciding on option (iii) above, as this could put the library or project team at risk of an infringement claim, resulting in a 'take-down' notice, or an embargo on the material and/or litigation in terms of the Copyright Act.
- Streamline copyright clearance procedures and reporting facilities to ensure ongoing control and efficient administration of copyright clearances. A standard letter of request is advisable. It can be modified according to specific needs. (See sample letter in Appendix 2.1).
- Digitise and store online the written permission for each work cleared. Include all written copyright permission letters or directives from rights holders with the individual digitised items/works, or create a separate 'permissions file' to keep a record of them on the system.
- Consider and record online the specific terms of permissions

granted, as well as any restrictions, embargoes or set periods of cover. If renewals are required in the future, the relevant dates will need to be available on the digital system, which should automatically generate reminders, when applicable.

- Consider all implications of embargoes and renegotiate with rights holders, when necessary.
- Request long-term permissions (i.e. in perpetuity) for backups, migration to new formats, preservation or replacement of damaged records, etc.
- Consider all conditions of born-digital licences and apply for permission for library activities not covered by the licences.
- Consider what to do if permissions are denied or if permission conditions are too prohibitive, as they might impact on accessibility or create gaps in the digital collection.
- Obtain permission to bypass or circumvent digital rights management systems/technological protection measures embedded in copyright works. These measures render works inaccessible or place restrictions on the works, e.g., geographic coding, no printing, scrambling devices, password protection, dedicated hardware, 'self-implosion' after a certain period, etc.
- Respond to and remove items in response to any 'take-down' instructions from rights holders who refuse to allow their works to be included in the digital collection, or who want them removed from the digital collection.
- Ensure, in the process of digitisation or digital curation, that the works are not altered, modified or have had sections cropped, damaged or deleted, to the extent that the moral rights of authors could be infringed.
- Set up an efficient accounting system for copyright clearances.

4.2 Budgeting for copyright permission

It is necessary for libraries engaged in creating digital collections and digital curation to have a plan of action concerning the copyright works that are to be digitised or the born-digital works that need to be migrated to newer technologies. An estimate from some of the larger publishers would assist in establishing the amount to budget for. However, this is a difficult task as copyright costs differ from country to country and currency to currency. Libraries need to decide on a suitable amount with which to commence their digital projects and then to budget more specifically in subsequent years when costs and procedures are more clear. Budgets also need to provide for renewals of copyright licences, depending on conditions given by rights holders, as well as for permissions for conversions to alternative formats for persons with sensory-disabilities, or migration to new technologies as well as the ongoing sustainability of collections, to ensure continuing copyright compliance. These costs are separate from those required for equipment, hardware and software, technology upgrades, personnel and administration tasks.

5. Protecting digital collections

The Copyright Act is vague on the protection of databases. "The requirement that causes most difficulties with database protection is whether a database is original enough to qualify for copyright protection. Originality in a database is usually a matter of how the information in the database is selected or arranged. But in many electronic databases the information is not arranged. It is dumped into the data-

base and a search program finds what the user wants. And because size is not an issue with an electronic database, some databases aim at including every piece of information about a subject. This can mean there is no selection of material. Databases of this sort are unlikely to qualify for copyright protection." (Hofman, 2009, pp.37-38)

There should not be any protection on public domain material included in the digital collection.

Copyright will subsist in any policy documents, manuals, guidelines, training modules, online tutorials, videos, CDs, DVDs, or any other works belonging to or generated for management of the digital collection.

Initial decisions need to be made concerning the following:
- The type of copyright works to be included;
- Ownership of copyright of the works to be included;
- The possibility of individual or shared ownership of included works;
- The possibility of the use of an open source licence for the digital collection, e.g. Creative Commons; and the type of licence to be used, e.g. Attribution Non-Commercial No Derivatives or another type;
- Responsibility for granting of copyright for the use of the digital collection by other parties;
- The conditions that will exist regarding the use or re-use of material from the digital collection;
- Whether the collection will be made accessible for non-commercial or educational purposes only;
- The copyright clearance procedures to be followed;
- The charges, if any, for administrative services and access;
- The management of charges;
- The measures to be taken to deal with infringers or users who default on payments, etc.;
- How to deal with embargoes required by rightsholders;
- Notices and disclaimers that should be provided to advise users about copyright in the digital collection.

.
If so desired, the digital project may use certain branding, logos or designs to identify itself. These can be registered under other Intellectual Property legislations, such as the Trade Marks Act or Designs Act.

6. Copyright administration

Libraries engaging in digital projects need to have dedicated personnel, preferably with knowledge of the copyright legislation and expertise to provide advice on issues of copyright, to assist with the application of copyright law and to attend to copyright clearances. Those members of staff would need to trace rights holders, send requests for permission, follow up requests, negotiate better terms, when necessary, and attend to permission renewals, embargoes and take-downs, as required by rights holders. They would also need to process invoices for payment of copyright fees and perform general administrative duties.

7. Intellectual Property Policy

It is important for the institution or library to have an up-to-date

Intellectual Property policy, that allows for the management of an the entire digital project. This policy should address all the items mentioned in sub-section 5 above, as well as other important issues relevant to the management of a digital project. The library should draft this policy with the assistance of a legal advisor or an Intellectual Property lawyer. The policy should be revised regularly to address any changes or new developments in the digital environment.

8. Notices/disclaimers

It is advisable to provide a number of notices or legal disclaimers on the digital project's database or website. These are:

a) **Privacy notice**
   To protect the privacy of users of the digital collection;
b) **Copyright disclaimers**
   To indemnify the digital collection and its website against any liability, in the event of incorrect or false information being contained in the digital collection, on its website, or other linked websites, where control of the content is not possible.

   To state that the opinions or views of those on the website or in the digital collection are those of the authors and do not necessarily reflect the views or opinions of the host institution.

c) **Notice of Invitation to Rights holders**
   When it has not been possible to trace rights holders of a work that has been included in the digital collection, it is necessary to place an invitation on the website requesting rights holders, or those who might know their whereabouts, to provide information about ownership of the works, so that correct clearance procedures can be followed and reasonable copyright fees can be negotiated, where applicable.

9. Conclusion

It is evident from this chapter that anyone embarking on creating a digitised collection will have to be rigorous in investigating ownership of works as well as applying for and obtaining copyright. The fact that the digitising process can change the original format of the work makes acquiring permission from the copyright holder absolutely essential.

References

Buys, R. (June 10, 2005) Personal e-communication to Denise.Nicholson@wits.ac.za *Copyright regulations on your website*.

Creative Commons. About Licenses. Available from: http://creativecommons.org/about/licenses (accessed 28 October 2009)

Dean, O. (1989). Handbook of South African Copyright Law. Cape Town: Juta.

Hofman, J. (2009). *Introducing Copyright: a plain language guide to copyright in the 21st century*. Vancouver: Commonwealth of Learning.

Hudson, E. and Kenyon, A. T. (2007). *Without Walls: Copyright Law and Digital Collections in Australian Cultural Institutions*. SCRIPTed. Available from: http://www.law.ed.ac.uk/ahrc/script-ed/vol4-2/kenyon.asp (accessed 28 October 2009)

Longe, O. B. (nd) *Intellectual Property Protection in the Age of Open Access and Digital Rights Management – Balancing the Odds*. Available from: http://www.ais.up.ac.za/digi/docs/longe_paper.pdf (accessed 28 October 2009)

Republic of South Africa. (1992). *South Africa: Copyright, Act (Consolidation)*, 20/06/1978 (1992), No. 98 (No. 125). Available from: http://www.wipo.int/clea/en/text_html.jsp?lang=en&id=4067 (accessed 28 October 2009)

Republic of South Africa (1978) *South Africa: Copyright, Regulations, 22/12/1978, No. R2530*. Available from: http://www.wipo.int/clea/en/text_html.jsp?lang=EN&id=4069 (accessed 28 October 2009)

Smith. A. (1995). *Copyright Companion*. Butterworths, Durban.

United States of America. (2008) *Glossary of Intellectual Terms*. Available from: http://www.america.gov/st/businessenglish/2008/January/20080109153644AKllennoCcM0.3526117.html (accessed 28 October 2009)

University of the Witwatersrand, Johannesburg. *Copyright Information*. Available from: http://web.wits.ac.za/Library/Services/COPYRIGHT.htm (accessed 28 October 2009)

Annexure 2.1

Sample letter of request for copyright permission

[*Date*]
[*Letterhead or Return address*]
[*Rights holder's name and address*]

Dear [*Sir or Madam*] [*Permissions Editor*] [*Personal name, if known*]:

[*Name of Institution and Digital Project Name*]

My institution/library is in the process of creating a digital collection under the Project name [*Name and description of project*]. I would like your permission to include the following material in this digital collection:
[*Provide full citation(s) with source information*]
The [*name of project*] will provide access to this material to [*describe most relevant user groups*] on an Open Access platform [*name - e.g. DSpace*].

If you do not own or control the copyright on the above-mentioned material, I would appreciate it if you could provide me with contact names and addresses of the relevant rights holder(s). If this is not applicable. I will accept that your permission confirms that you hold the right to grant the permission requested here.

Permission requested includes non-exclusive world rights in all languages to use the material and will not limit any future publications, including future editions and revisions by you or others authorised by you. Permission is also requested to allow our project to migrate our digital collections, including the abovementioned works/material, to new technologies as and when necessary, for the purposes of preservation and long-term digital curation.

I would greatly appreciate your written consent to my request at your earliest convenience. If you require any additional information, please do not hesitate to contact me. I can be reached at: [*Your full contact information*]

Your assistance in this matter will be much appreciated.

Thank you in anticipation.

Yours faithfully

[*Name, Position and Signature*]

(An adapted version of the Copyright Management Centers' General Permission -Model at http://www.copyright.iupui.edu/pgeneral.htm)

Chapter 3

Collection development

1. Introduction

Since the creation and establishment of the library of Alexandria in the Third Century BC, there has always been a vision of a library without limits, a universal library aspiring to contain all known information. With the rapid and exponential accumulation of information on the Internet, and with ever-widening access to the World Wide Web by users across the globe, that idea no longer seems totally far-fetched.

Collections housed at different institutions, can be linked on the Internet to form one complete collection on a certain topic or source. Therefore it needs to be well-managed with clear collection development guidelines and policies.

2. Collection development

Collection development is a broad description for the management of collections of information resources and involves their identification, selection, acquisition, evaluation, and sustaining.

Collection development is normally connected to:
- The core functions of the housing institution;
- Means of access;
- Usefulness;
- Relevance;
- Available storage space;
- Diversity of user information needs;
- Financial constraints.

Every digital collection is unique with its own users, goals and needs.

Good digital collection development is an active and ongoing process between archives, libraries, museums as well as between collectors and users of information. While a digital collection does not necessarily need to be housed at one place, it can be accessed from various institutions beginning at one point - the Internet.

Long-term digital stewardship is one of the most important aspects to be addressed in collection development. It is essential that the level of the institution's contributions to the longevity and usefulness of the digital content, within or apart from a formal digital preservation programme, should be clearly established from the outset. The sustainability of the collection plays an important part in the success and value of the collection as well as future access to it.

3. Digital collections

Digital collection development can be seen as part of the broader concept of collection development and is linked to the vision, mission and goals of the housing institution or organisation.

Figure 3.1 How different types of material can be part of a single digital collection.

All types of material can be part of a digital collection.
Material housed at various destinations can form one complete digital collection

A digital collection can be described as an organised group of objects, managed and displayed in such a way that it is meaningful and usable with the user when accessed.

In the digital realm there are no absolute rules for creating good digital collections. Each collection varies by type, sources included and relationships to other formats and collections.

There are important distinctions between collections of born-digital material and collections created through digitising material in different media or formats such as paper, film, video or even three dimensional objects.

A digital collection can consist of a variety of format types, depending on the type of information data and composition of material in a collection.

4. Measures to identify a 'quality' digital collection

The NISO Framework of Guidance for Building Good Digital Collections[21] identifies nine principles for the identification of a good collection.

A digital collection
- is created according to an explicit development policy;
- has well-composed metadata that describes the collection, scope formats, usage, authenticity, integrity and technical aspects for usage and storage;
- should be managed during its entire life cycle, and preserved with well-established, but fluid preservation policies;
- should include measures to provide availability to disabled people;
- must respect intellectual property rights;
- must adhere to international standards;
- should be interoperable (usable by different systems and applications);
- must integrate with users' workflow;
- must be sustained and accessible.

A good digital collection is developed around clearly stated policy guidelines and principles, that stipulate what should be collected and preserved.

5. Selection criteria

The selection of the material in a digital collection should receive careful consideration. Questions to ask are:
- What is the core function of a collection? This means that valuable assets that a library wishes to make available should be the focus.
- Who is the target audience? How is the intentional use of the collection envisaged: internal, limited external or, open access on the World Wide Web.
- What aspect of the documentation will be collected? Will it be only the original (master) or derived copies, linked pages or only the links, only the final file derived from the original?
- Will the relationship of digitised documents to an original source, form an integral part of the collection, or will there only be references in the metadata?
- How should copyright clearance for the physical material be obtained? What will the budgetary implications be for the host organisation?
- Who will take responsibility for sustaining the collection and for undertaking regular risk assessments, including checking the on-going functionality of the documents?
- Who will be responsible for the migration and/or emulation or any future activities that are essential for ensuring long-term access and preservation of the collection?

Digital collection development must include the economic aspects for the establishment and management of collection content during its anticipated lifespan.

6. Developing the collection policy

The success of the criteria for digital collections can be measured in terms of cost and /or value, sustainability and trust (reliability). *Good* means that a collection is interoperable, reusable, persistent, can be verified, and contains documentation to support its authenticity and intellectual property rights. A digital collection is fragile and expensive to maintain. As a result its future consistent availability is a critical consideration in the decision to start a specific collection.

It is also worth knowing whether similar collections exist elsewhere and whether there is the possibility that collaboration and shared responsibility exists.

Therefore, policy developers should do the following:
- Keep the core business functions of the host institution in mind when considering selection criteria;
- Stipulate the possible uses of the collection, identify the target audience and consider possible unexpected users;

---

[21] NISO Framework Working Group, A Framework of Guidance for Building Good Digital Collections, 3rd Edition, 2007

- Address the preservation and performance needs of the collection objects, as well as access requirements for users;
- Include measurements for the electronic objects to be readable and interpretable on a variety of computer systems i.e. interoperability;
- Describe the ongoing management of the collection, staff involvement, their various specific roles and workflow;
- Address the budget for sustaining the collection.

7. Description of the collection

Effective preservation methods start with classification principles. The metadata of an object can be viewed within its context, as a building block that others can reuse, repackage, repurpose and build services upon. Descriptive metadata give meaning to the electronic object and provide for its search ability and access.

Metadata schemes support the conservation methods applicable to a digital collection. More than one metadata schema can be used in a collection, as the format types of the objects may differ and it is not always possible to describe this with a single metadata scheme.

8. Construction of a digital collection

A digital collection can consist of the intellectual output generated by the staff of an institution during the course of their work. This can include emails, text, datasets, presentations, and so on. Other material can include research output, like theses, dissertations and articles, donated material collected by individuals, as well as material harvested by the institution. Because the size and formats of these objects can vary, different format types for digital objects need to be addressed separately.

The International Federation of Library Associations (IFLA)[22] document for the descriptions, theory and practice of collections provides valuable information on standards for collections and ingest in repositories and databases.

Typical file formats are:
- email (ASCII, .txt);
- text files (.doc and other);
- music (MP3, WAV);
- pictures (TIFF, Jpeg);
- videos (AVI, MP4, flash);
- history (various formats, i.e. doc/JPEG);
- datasets (various formats; sometimes proprietary datasets customised especially for a datum type, e.g. doc/xls/png/gif).

9. Collection development and metadata[23]

The use of metadata to describe a collection should form part of the collection policy. Metadata can explain the content of a collection so

that a user can discover its characteristics, including scope, format, restrictions on access, ownership, and any other information that would be significant for determining the collection's authenticity, integrity and interpretation[24] (NISO – Collections Principle 2).

Two approaches to the storage of metadata associated with a digital object contained in a collection, are:
- Metadata stored internally with the object;
- Metadata held externally, with references in the metadata to the digital object.

The contents of a digital collection and its objects should provide for sufficient detail for potential users to assess whether it is suitable for their needs. The metadata should include, where possible and applicable:
- Title;
- Author (creator);
- Period of creation;
- Name of country of origin or residence of the object;
- Language of the display/creation;
- Description of the contents.

The above metadata can also be searched and retrieved by search engines and serve as points of access.

The provenance of a collection needs to be documented and must accompany the collection in the form of metadata. The information should include the method, reason and persons responsible for the creation of the collection.

Automated metadata tools support formal preservation metadata elements and have surfaced in a variety of digital asset management environments, for example PDF-documents or image files like TIFF.

Awareness of preservation metadata (see PREMIS[25]) and the practical applications thereof, is important during the development phase of a collection. The metadata should describe the type of object and its format. Multiple collections can be combined with links and metadata for search and preservation purposes, and displayed to the user as a single collection. These links should be maintained. It is best to use link resolvers for this purpose.

10. Curation of digital collections

10.1 Archival Record

The definition of an archival record as stated on the website of the National Archives of South Africa reads as follows:

'... the 'archival record' came to mean recorded information, without regard to form or medium. Essential to this definition is the understanding that a record takes on archival quality if the information it contains has enduring value. Archival records can therefore be paper-based textual records, electronic records, audio-visual, photographic or cartographic material.'[27]

A digital collection should be supported by all necessary, implementable methods to ensure the ongoing retrieval and exploitation of the information resource.

22 www.ifla.org/IV/ifla73/papers/125-Hakala_Keskitalo-en.pdf
23 See Chapter 6: Metadata
24 NISO Framework Working Group, *A Framework of Guidance for Building Good Digital Collections*, 3rd Edition, 2007
25 http://www.loc.gov/standards/premis/
26 www.niso.org/workrooms/mi/Z39-91-DSFTU.pdf
27 http://www.national.archives.gov.za/

Figure 3.2   The NISO Z39.91[26] standard is used to define a means of describing digital collections.

Derived from http://www.loc.gov/standards/premis/understandingpremis.pdf by Priscilla Caplan for the Library of Congress: *An overview of the PREMIS Data Dictionary for Preservation of Metadata*

Images: © istockphoto.com/Art-Y, duncan1890

10.2 Risk Assessment

The curators of a digital collection should understand the basic technologies and metadata for the preservation and risk assessment of the collection. A long-term measurement plan, for quality assessment and survival of the objects, supported by computer systems and protocols enabling a well managed digital collection, should form part of the policy. The policy can be revised when new technologies emerge.

The International Organisation for Standardisation and the International Electrotechnical Commission (ISO/IEC) 27002-2005 standard can be used in the risk assessment of digital collections and its objects :

'ISO/IEC 27002:2005 establishes guidelines and general principles for initiating, implementing, maintaining, and improving information security management in an organisation. The objectives outlined provide general guidance on the commonly accepted goals of information security management.'

10.3 Digital Object Lifecycle

A digital object is a small part of a digital collection in need of preservation principles for future sustainability.

The Digital Curation Centre (DCC) defines digital curation as:
'... maintaining and adding value to a trusted body of digital information for current and future use ...'[28]

The Digital Curation Centre has developed a lifecycle model for digital object management which can be obtained from their website: http://www.dcc.ac.uk/diffuse/.

10.4 Born-digital Material

Different challenges arise when the collection involves 'born-digital' materials that should be preserved. These include the following:

- Born-digital materials can have 'distributed' (NISO[29]) ownership, i.e. emails, Web pages, blogs and wikis.
- It can be difficult to convince owners to donate their material to repositories, archives, museums or libraries.
- The material is not always well-documented with metadata and needs trained staff to dismantle the digital content donated to the institution for preservation.

The Open Archival Information System (OAIS)[30] is used by people and systems to preserve information and make it readily available. The model provides a framework for increased awareness and understanding of digital archival concepts that are needed to store file formats in digital collections. The lifecycle of a file format is influenced by changing technologies, changes in targeted audiences, or change in the way information is accessed and used.

10.5 Technology Obsolescence

The obsolescence of hardware and software is well known in the digital realm. Many users have experienced the changes from floppy-to-CD/DVD-to-flashdisk usage, as well as known file formats that are no longer in use by their systems. Steps can be taken to ensure that stored information remains available even though the original technology has become obsolete. Tested methods for digital curation include the following:

---

[28] http://www.dcc.ac.uk/about/
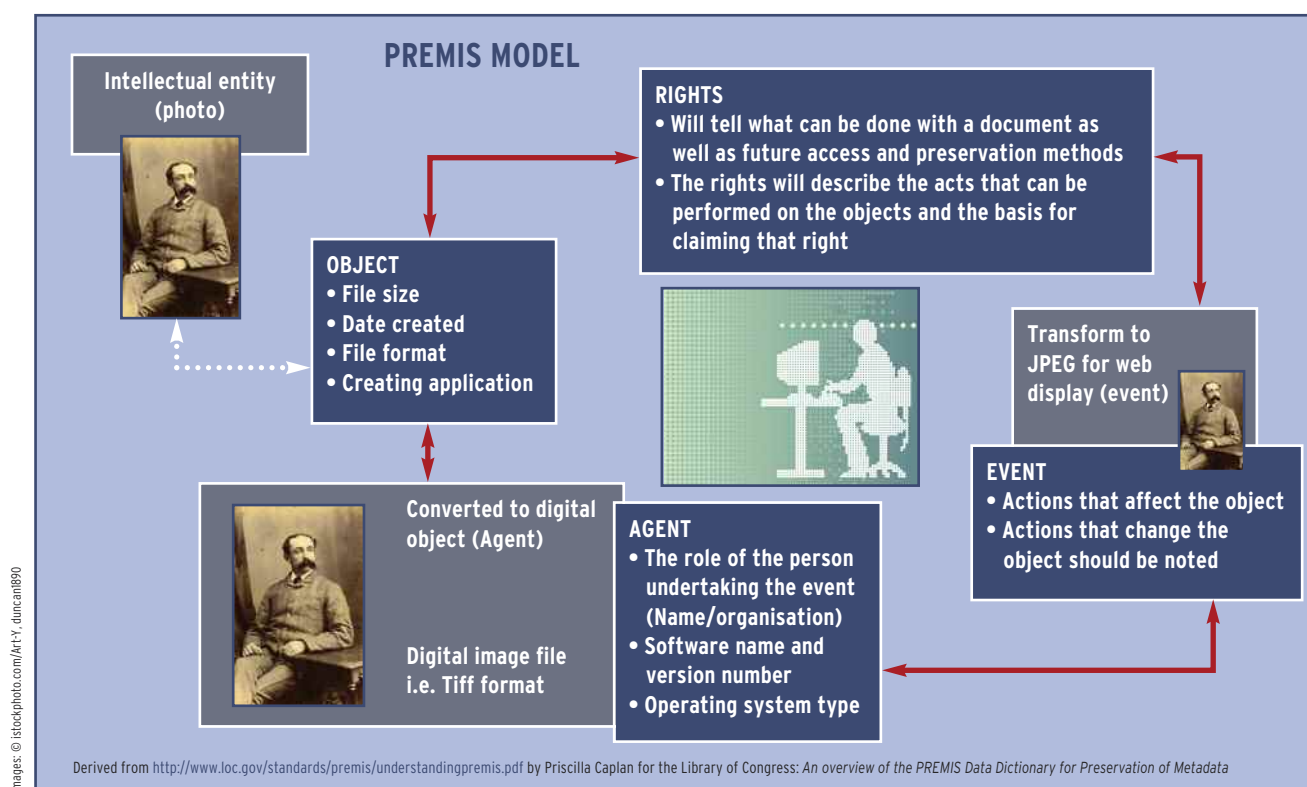[29] NISO Framework Working Group, *A Framework of Guidance for Building Good Digital Collections*, 3rd Edition, 2007
[30] public.ccsds.org/publications/archive/650x0b1.pdf

- Migration[31,32] which involves the movement of information from one format to another, e.g. text in WordPerfect moved to MSWord and, where necessary, the renewal of the format version;
- Emulation, which involves the building of emulation hardware and software that emulate the old systems and software used in the creation of a particular digital format. Emulation makes it possible for a digital format to be visible in the original form in which it was created. This method is expensive but is currently widely seen as the solution to possible obsolescence problems.

The change in the nature and usage of information has transformed organisations such as libraries and museums. The management and sustainability of collections goes far beyond paper preservation.

11. Availability of a collection

The availability, usability and accessibility of a digital collection differ from an analogue collection inasmuch as it is not bound to the place of storage. While, digital information can be restricted and can only be accessible from the local area network of a host organisation, it can also be widely accessible through the World Wide Web.

The planning of a digital collection should include its fitness for use by the targeted audience. The type of information contained in a collection will define its accessibility. The guidelines for the development and creation of a digital collection should include the following considerations.

- Disabled people are often overlooked in the development phase of a digital collection. The South African Constitution not only provides a right of access to information, but it also protects the rights of disabled persons. Therefore the collection manager is legally obliged to include all possible methods for making the information available to all. This includes voice recognition for visually impaired persons, cursor movement and sound description for deaf people and so on.
- Access will be influenced by the number of potential users who will be capable of using the technology required for access. The planning phase of collection development should include an indication of the number and categories of users who may find the technology a barrier to access.
- Ease of use and file format types included in the collection will have an influence on access. Software used in the creation of a digital object might not be interpretable by all computer operating systems, therefore it is best to convert a display document to a 'uniform format', such as PDF, for interoperability.
- Bandwidth (or the lack thereof) can be a major problem. The minimum browser version and bandwidth requirements should be taken into consideration when digital objects are manipulated for presentation.
- Web pages should be tested against the various available Web browsers as the display can differ and result in wrongly displayed information.
- The colour settings of a monitor screen can be manipulated by the user. Where colour is essential, it might be necessary to display a colour chart against the digital object for correct interpretation.

- Different screen resolutions can influence the size of the display. It is advisable to inform the user of the preferred settings.
- Diacritical characters can also be problematic, and can be displayed wrongly on some browsers. A note at the home page of a digital collection should inform the users of the preferred browser for display.
- The digitisation process can result in the loss of recognisable text. However, this can be rectified by the recognition of the text by OCR software.
- Language barriers in digital collections should be addressed. If the text is not translated, the metadata should contain enough information for the user to be able to interpret the content of an object.
- Redaction of offensive words in text might be considered but should be applied consistently.

12. Intellectual property rights

The right to access a collection should be cleared, noted and respected. The intellectual copyright of digital objects must be retained in an object and embedded with metadata. For born-digital material it is best done by the creator of a digital object. When material is donated to an institution the following questions must be answered.

- Can the material be migrated?
- What permission is given for the use of the content?
- If it is still in physical format, can the material be digitised?

The idea of a '*Digital Will*' is becoming popular worldwide. Such a will stipulates the use of donated digital material in a collection.

Good record keeping of rights holders and embargos placed on collections is an absolute necessity. Failure to do so can expose the custodian to legal risk. The unique identification number of a document should link the rights given to the actual document.

The fair use of information as well as its research and teaching value, should be taken into consideration where the digital ownership and rights are not traceable.

Where the 'opt-out' policy is used a note should accompany the object to state that the institution can be notified when bridging copyright and that it will subsequently remove the object from the presentation collection.

Example of a Copyright Notice by the South African History Archive (SAHA)[33]

Copyright Notice: '… Please note: Every attempt has been made to obtain copyright clearance for the material on this website. However, we have not been able to trace some rights owners. We have included their material in this digital collection, but invite anyone who can assist us with contact details for the rights owners, to contact us at … … so that we can obtain formal permission'

Intellectual property rights are internationally protected by treaties. International treaties provide the minimum requirements to which all signatory states must adhere. Examples of such IP treaties include the following:
- **WIPO**[34]
WIPO was created in 1967 and is responsible for the promotion of the protection of intellectual property throughout the world
- **Berne Convention (Paris 1971)**[35]

---

31 http://www.kb.nl/hrd/dd/dd_projecten/projecten_migratie-en.html
32 http://en.wikipedia.org/wiki/Data_migration
33 http://www.saha.org.za/

The Berne Convention protects the copyright of authors and artistic works. It was first established in 1886 and has been revised many times since then.

13. Collection development and the influence on the workflow of staff

The preservation and management of a digital collection has a direct influence on the workflow of the staff in an organisation. Old ideas of curation must be transformed and refocused. Each object in a digital collection should be regularly quality-checked and basic metadata should be added to describe and accompany the object.

The trust of the host organisation must never be in doubt and staff must be trained to retain and uphold it. Staff should also be trained in the use of different file formats and know where to store the information until it is ingested in a collection, repository database, or other method of digital archival storage.

Quality checks on file formats for possible data loss, maximum information retrieval, as well as the on-going stability of the storage containers should be done by personnel. Databases containing the information should be updated regularly with information and dates of quality assurance checks.

The archiving of digital collections needs dedicated and constant attention. Staff should be trained in the understanding of AIPs (Archival Information Packages) which include metadata and preservation methods, storage of the data, as well as the care and custody of fragile digital collections. The digital environment is fluid and ever-changing, therefore training should be done on a regular basis to ensure a stable and well-preserved collection.

Management of digital content is now seen as part of the basic services of an organisation. Staff should be encouraged to communicate across departmental boundaries to achieve greater exposure to available expertise.

In-house training, attendance of conferences, registration on email list serves, the use of RSS-feeds and regular searches on the Internet for information regarding sustainable and preservable digital collections is recommended.

14. Authenticity

The authenticity of a digital collection must never be in doubt. The name of the custodian organisation is at stake. All known measures must be taken to ensure an objects' authenticity. Metadata that accompany the collection and its individual objects as proof of authentication must be as comprehensive as possible.

Written statements should be attached to the collection to enhance proof of authenticity.

While comprehensive documentation about a collection, work done on a collection during its development, and steps taken to stabilise or normalise it, involves a significant amount of time from the staff, this will ensure its ongoing, long-term validity.

[34] http://www.wipo.int/portal/index.html.en
[35] http://www.law.cornell.edu/treaties/berne/overview.html

15. Conclusion

Many different tasks are involved in the identification and acquisition of material as well as in the sustaining of a digital collection. Collection development is associated with the core functions of the housing institution and as such is influenced by benefits and barriers of the institution. It is of utmost importance to develop a collection policy that will adhere to the policies as well as the financial and human resources of the institution to ensure that the selection, storage and retrieval of material can remain constant and authentic.

References

Consultative Committee for Space Data Systems. (2002). *Reference Model for an Open Archival Information System (OAIS)*. Available from: public.ccsds.org/publications/archive/650x0b1.pdf (accessed 7 April 2010)

Cornell University Law School.(nd). *Berne Convention for the Protection of Literary and Artistic Works* (Paris Text 1971). Available from: http://www.law.cornell.edu/treaties/berne/overview.html (accessed 7 April 2010)

Digital Curation Centre. (nd) DCC DIFFUSE *Standards Frameworks*. Available from: http://www.dcc.ac.uk/diffuse/ (accessed 7 April 2010)

Hakala, J. and Keskitalo, E. (2007) *Description of collections in Theory and Practice*, Paper read at the 73rd IFLA General Conference and Council of the World Library and Information Congress, 19-23 August 2007, Durban, South Africa. Available from: www.ifla.org/IV/ifla73/papers/125-Hakala_Keskitalo-en.pdf. (accessed 7 April 2010)

International Organization for Standardization (ISO), ISO/IEC27002:2005. (2005). *Code of practice for information security management*. Available from: http://www.iso.org/iso/catalogue_detail?csnumber=50297 (accessed 7 April 2010)

National Archives and Records Service (NARS) Available from: http://www.national.archives.gov.za/aboutnasa_content.html (accessed 7 April 2010)

NISO Framework Working Group. 2007. *A Framework of Guidance for Building Good Digital Collections, 3rd Edition*, Available from: http://framework.niso.org/ (accessed 7 April 2010).

NISO Z39.91-200x. (nd) *Collection Description Specification* Available from: http://www.niso.org/workrooms/mi/Z39-91-DSFTU.pdf (accessed 7 April 2010)

PREMIS Editorial Committee. (2008). *PREMIS Data Dictionary for Preservation Metadata version 2.0, 2008* Available from: http://www.loc.gov/standards/premis/ (accessed 7 April 2010).

South African History Archive. Available from: http://www.saha.org.za/ (accessed 7 April 2010)

Townsend, S. Chappel, C., Struijvé, O. (1999) AHDS Guides to Good Practice. *Digitising History: A Guide to Creating Digital Resources from Historical Documents*. Available from: http://hds.essex.ac.uk/g2gp/digitising_history/sect51.asp (accessed 7 April 2010)

World Intellectual Property Organization (WIPO) Available from: http://www.wipo.int/portal/index.html (accessed 7 April 2010)

Chapter 4

Objects

1. Introduction

This chapter will explain the term 'object' in the digital environment. The formats that act as a carrier for a digital object will be explained as well as the interpretation and uses of the formats. A well-managed object should be exchangeable across different platforms, be widely accessible, and formatted according to recognised international standards for future use and preservation. The use of different file formats are unique to its content and intended use, therefore the intended audience should be taken into consideration in the creation of a digital object.

2. Explaining the term object

A digital object can be defined as any item that can individually be selected and manipulated; it is a self contained entity that represents digital information consisting of data and procedures to manipulate the data (images, text, videos, etc).

An object can be described as a unit of information, or an item stored in a digital format that consist of electronic data and metadata associated with the object. The electronic file 'carrier' for the object is called a format. Different formats are used for the interpretation of the bits and bytes of a digital object, depending on the type of information it contains and its intended use. Object types can be text, photos, 3D-objects, sound or any other multi-media format.

The data which forms the object is basically a combination of a binary string of zeros and ones and can be interpreted by a computer or electronic device and displayed for viewing by users of these devices.

Figure 4.1 A binary string

$$0101110010O111$$

An object can consist of several types of digital material and the size of the object depends on the amount of data contained in the data file. The formatting software used in the creation of the object retains the intended structure and behaviour of the object, and gives meaning to the binary values.

A digital object can be 'born-digital', meaning that it was originally created on an electronic device, or converted from an analogue source to a digital format. Most, if not all, of the information needed to manage a digital object throughout the whole of its lifecycle is captured in the administrative metadata (see Chapter 5: Metadata).

The digital object can be a 'stand-alone' entity or it can consist of several types of digital material.

Figure 4.2 Example of a single digital object that has no relationship embedded
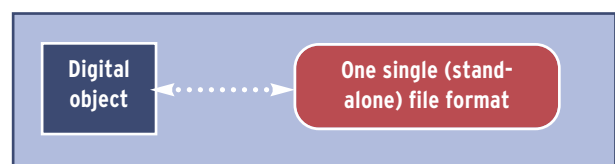

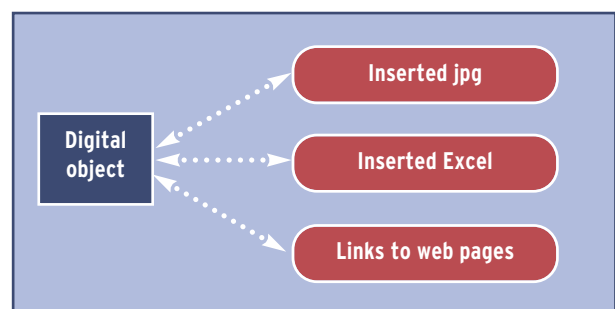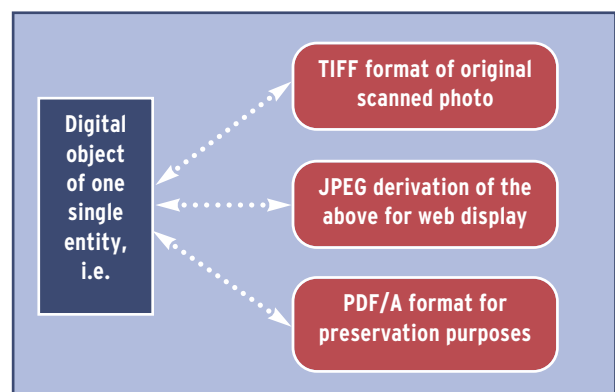**Figure 4.3 Example of a digital object with multiple formats embedded to form one entity**

Figure 4.4 Example of a single-source displayed in multiple format types

Characteristics of a well managed digital object are:
- They are exchangeable across different platforms (i.e. interoperable).
- They are interpretable by all kinds of electronic devices.
- They are searchable for retrieval and access.
- They are created and stored according to established international standards for future use and preservation.

The National Information Standards Organization (NISO) recommends six principles to identify a good object.[36]

A good object:
- is in a format that supports its intended current and future use;
- is preservable;
- is meaningful and useful outside of its local context;
- will be named with a persistent, globally unique identifier that can be resolved to the current address of the object;
- can be authenticated;
- has associated metadata.

3. Formats

The stored format of a digital object is established by the application used to create the object. File formats of the software that interpret the digital object are defined by an extension at the end of the filename and start with a .(dot) followed by a series of three or four letters (e.g., *photo.jpg, filename.docx*, etc.). File formats identify the type of file and allows the user to acknowledge the program used for editing of the object.

The final file format of an object should be broadly accessible across platforms, and based on open standards.

There is a difference between file formats and codecs. In the case of audio files, the codec is the software program that compresses or decompresses the digital audio data according to the file format specifications. [37]

A good object can be preserved for future access and should be able to overcome changing technologies and software versions. Best practice is to standardise on file formats specified for specific kinds of digital objects. Formats that are widely used and have published specifications are most likely to survive migration (open standard). The Tagged Image File Format (TIFF), an uncompressed lossless format (no data is lost when viewing and saving the format and usually results in the generation of large file sizes) is one such format and is widely

used in the digitisation process for archival images. The format carrier represents the structure of the object data as well as its display, if the reader of the format is available.

Master digital objects intended for archival purposes (or to have a long life span) should be created with digital preservation in mind and are therefore normally saved in a lossless format, with no optimisation or compression to the file during the editing and saving process. Uncompressed files comprise large file sizes and should be stored with accompanying metadata to describe the file content. Non-proprietary formats that do not contain patented technologies are the preferred choice for objects that need to be preserved for long-term use. Creators of digital objects normally know what the object will be used for, and what the intended life span of the object will be.

3.1 Content category of digital object – object types

Digital objects can be categorised into two types, namely raster (composed of pixels) and vector (composed of paths).

a) Bit-map (raster) images use a grid or matrix, where each element has a unique location and independent colour value, known as pixels, to represent images. Images created from a scanning device are raster or bit-map images.
For a detailed explanation of raster imaging, please refer to the 2004 document of the National Archives and Records Administration: *Technical Guidelines for Digitizing Archival Materials for Electronic Access: Creation of Production Master Files – Raster Images* (Available from: http://www.archives.gov/preservation/technical /guidelines.html).
b) Vector graphics are lines and curves and are defined by mathematical objects called vectors. Vectors describe an image according to its geometric characteristics. The object can be moved, resized or colour-changed without losing any quality to the graphic. Text documents and computer line art are examples of vector-based objects. Some other well-known formats in which the boundaries between raster and vector images become less obvious are JPEG2000, Flash, SWF and SVG.
c) Metafile formats are able to encapsulate both raster and vector images, i.e. both text and graphics. The contents of these formats use different operating systems and software for their display. Examples are the Computer Graphics Metafile (CGM) which will run on most computer operating systems with PostScript (PS) that are designed to provide detailed instructions to printers with Page Description Languages (PDLs).

36   NISO framework working group, et al., 'A Framework of Guidance for Building Good Digital Collections', National Information Standards Organization (NISO), http://www.niso.org/publications/rp/framework3.pdf (accessed January 15, 2010).
37   Webopedia, http://webopedia.com/DidYouKnow/Computer_Science/2005/digital_audio_formats.asp. (accessed January 15, 2010).

The Portable Document Format (PDF) is a metafile based on PostScript, but adds functionality and supports a range of compressions, both lossy and lossless. The PDF format is increasingly being used as a container format for exchanging digital documents because of its interoperability. It is also known as a Uniform Format.

3.2 Preservation and display formats

The use of different file formats is unique to its content and intended use. For Web display the bandwidth needs to by taken into consideration, it is therefore desirable to compress an original lossless master to a smaller download size for screen display. Although the difference between the two object types do not seem to be large when viewed on a monitor, the download size of a compressed image results in a smaller file size which is more user-friendly. The larger lossless format needs to be kept and stored separately for preservation. (See also Chapter 7: Digital Preservation)

Image stored in lossless Tiff format

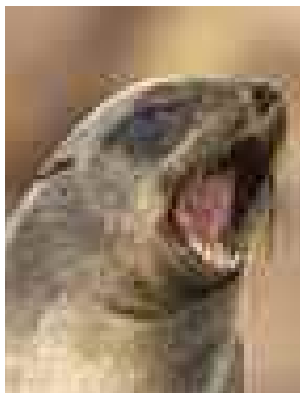**The same image compressed to a lossy JPEG format for screen display**

Image © Loretta Steyn

3.3 Different format types and their use in digital documents

a) Textual (i.e. Microsoft Word .doc. and docx)

Digital textual documents can be created by different word processing software which store the format structure as well as the typed information i.e. MS-Word. For cross-platform exchange the Rich Text Format (rtf) was developed by Microsoft, but this format has limited use. This format does not store the structural information of the document and results in a small file size. The TxtFormat (txt) is a Macromedia Flash v2 component for formatting text-field contents at runtime.

Text created by word processing software and intended for exchange is often converted to the PDF format to ensure interoperability. The Reader for PDF documents is freely available on the Internet and can be used across different platforms. The PDF format was officially released as an open standard in 2008 (ISO/IEC 32000-1:2008). The PDF/A format is widely embraced for archival purposes and can also be seen as a method to 'freeze' a document in its original state.

Standard practice in digital imaging:

The scanned format for master textual imaging is TIFF 300-600 dpi colour or grayscale depending on the original. Spatial resolutions should be based on the size of the text found in the document. The derivative is normally displayed in JPEG or PDF-formats. Optical character recognition (OCR) is recommended to make the text machine readable and full text searchable.

Access to textual material can also be enhanced through SGML/XML markup schemes such as the Text Encoding Initiative (TEI). (See Chapter 5: Metadata)

b) Photographs - (i.e. Camera RAW, TIFF, JPEG)

The default storage format set for most digital cameras is JPEG which is an acronym for the name of the committee that created the standard, the Joint Photographic Experts Group. Images displayed in JPEG format are widely used for the presentation of images on the Web.

Hint: The 'point and shoot' digital camera systems, mostly used for capturing digital photos is not suitable for digitisation projects as the lens quality is limited and no studio flash synchronisation is available.

The standard file format that is used for master document imaging is TIFF. The TIFF format was originally created to get scanner vendors to agree on a standard file format for scanned images. TIFF files are lossless files comprising large file sizes which are often not suitable for display on the Web because of its size. TIFF is regarded as a good archival format and used for preservation purposes.

Hint: Consumer cameras rarely allow for Adobe RGB (red, green and blue) colour space or the RAW file format (minimally processed data and not ready for printer use) which is needed for original digital master images.

The Portable Network Graphics format or PNG-format, utilizes lossless compression and was originally developed to replace the older GIF-format. PNG has many advantages for Web designers, but unfortunately it is not yet supported by all browsers. The format can be used successfully as a derivative file format for scanned images, and then converted into a PDF file for optimisation and quick Web view.

The Joint Photographic Experts Group File Interchange Format (JPEG2000) is a wavelet-based standard for the compression of still digital images. Although JPEG2000 still use lossy compression, it may have potential for becoming the file format of choice for archival master images in the future.

Hint: When digitising original photo captures, it is recommended that digitisation of photographs be from the negative or the earliest generation of the photograph and stored in TIFF. However, when the photograph was developed according to the specification of the artist him or her self, the developed photograph should be digitised.

The master image should be digitised in full colour range in order to create a more accurate image as a digital master should be rich enough to accommodate future needs and applications. Where a

significant amount of information exist on the back of the photo, it should be digitised as a separate image file.

Standard practice in digital imaging:
The scanned format for photographic imaging is TIFF 300-600 dpi colour or grayscale depending on the original. The derivative is normally displayed in JPEG but can also be embedded in other file formats such as PDF. For authentication, digital photographic editing is not recommended for master digital images.

c) Graphics (i.e. engraving, lithography, line art, graphs, diagrams, illustrations, technical drawings and other visual representations)
The originals of digital graphic objects are very often created in vector or metafile formats and later converted to image display formats such as TIFF or JPEG. For digitisation it is important to note that graphics are mostly two dimensional and the master image should be scanned in full colour, saved in a lossless format such as TIFF, with the derivative in JPEG format.

d) Datasets (i.e. Microsoft Excel(.xls), or any unknown file format created for a data set)
A dataset can be defined as any organised collection of data or information that has a common theme. A dataset might be a list of objects, a digital map, records of geological borehole samples, a collection of photographs of a certain subject, etc. One may also see datasets as individual-level results of a survey, conceptualised as a table or 'matrix' where the rows are individual respondents and the answers given by each respondent or values derived from those answers. Data sets can consist of different digital data information types and can then be embedded into one document, i.e. a spreadsheet, which should be normalised for interoperability into a metafile format.

Standard practice in digital imaging:
The standard practice for digital imaging of text and images applies for datasets. It is recommended that links be built to the different data with XML or in a PDF-file in order to retain the complete set. For preservation of datasets it is recommended that it should be accompanied by a comprehensive records-management program and formally established records procedures.

Additional reading recommended: http://hds.essex.ac.uk/g2gp/digitising_history/index.asp

e) Harvested websites (HTML)
Web crawlers harvest websites for storing of older webpage versions and later web pages for archival purposes. Snapshots of websites and/or complete websites, with the navigation retained, can thus be archived. A well-known web harvester initiative is the Internet Archive[38]. During a web archiving activity the web page with its embedded multi-media and linking structure are collected. Most harvesting use ARC and WARC formats both designed for web archiving. ARC was developed by the Internet Archive to support its work while the WARC format (based on ARC) is regarded as a

container that permits one file to carry a very large number of constituent data objects, of unrestricted type, for the purpose of storage, management and exchange. WARC is thus regarded as the preservation file format generated for use by the Heritrix web crawler and was published as an International Standard in June 2009, namely ISO 28500:2009.

The Web Curator Tool (WCT)[39] manages selective web harvesting and is designed for use in libraries and other collecting organisations. These sites are mostly tagged in XML for archiving purposes and need special metadata inscriptions for best archival practices.

Additional recommended reading: http://www.loc.gov/webcapture/technical.html

f) Professional CAD-CAM, engineering, manufacturing applications
The term CAD/CAM is an abbreviation for Computer-Aided Design (CAD) and Computer-Aided Manufacturing (CAM). These file formats cannot be read without their viewers but can be embedded into carrier files such as a dataset.

g) Geospatial, GIS
The term geospatial (used in conjunction with geographic information systems and geomatics) is used to describe the combination of spatial software and analytical methods with terrestrial or geographic datasets[40], whereas Geospatial Information Systems capture, store, analyse, manage, and present data that are linked to a location.[41]

In addition, frozen GIS data can be output as a bit-mapped image but then the structural information is lost. These objects are mostly born-digital, vector-based and marked up in GML (Geography Markup Language) and the Spatial Data Transfer Standard (SDTS).

h) Digital audio (i.e. MP3, WAV)
Digital audio refers to the transmission of sound stored in a digital format on a wide variety of storage devices, including the computer itself. Although many formats exist for recording and playing digital sound, many of these file formats are software dependant. Software conversion tools can be downloaded from the Internet to convert one format type to another.

Three main audio categories of format types exist for digital sound:
- uncompressed – all the data is available (i.e. WAV)
- lossless compression – although compression is applied, an exact digital duplicate of the original audio stream is created during compression, with no irreversible changes from the original version during playback.
- lossy compression – some loss of data as redundant or unnecessary information is removed by the compression algorithm (i.e. MP3 and Real Audio)

The Waveform Audio Format (WAV) is an audio file format standard for storing audio bitstreams such as sound and music on computers

---

38 The Internet Archive: http://www.archive.org/. (accessed 15January, 2010)
39 Web Curator Tool: http://webcurator.sourceforge.net/. (accessed 15January, 2010)
40 For a definition and explanation, see the Wikipedia reference: http://en.wikipedia.org/wiki/Geospatial
41 For a definition and explanation, see the Wikipedia reference: http://en.wikipedia.org/wiki/Geographic_information_system

and is widely used in the digitisation process as the archival storage format.

MPEG3 and Real Audio formats are used for the Web presentation of digital converted sound with a growing interest in the Ogg (.ogg) format which is open standard and uses the Vorbis audio compression scheme. MPEG has also created open standards, but one format may not necessarily be compatible with or based on the same MPEG standard, because of proprietary components.

i) Music scores

Music can also be digitally created by a variety of available software such as Fort Notation Software which is freely available on the Internet[42] and can be exported to JPEG and MP3 formats.

The digitisation method for sheet music is the same as for text. The derivative is binarised (converted to black and white thresholds) for recognition of the music score.

A 99% accuracy rate can be reached with a clean scan during optical music recognition. By OCR-ing of text, the words are parsed sequentially while music notation involves parallel elements. Expression marks, dynamics and spatial relationship amongst others, are important for the expression of music. But due to number of variables involved in the expression of music, interpretation problems can occur in scanned music notations.

An example of proprietary optical music recognition (OMR) software is PhotoScore An example of open source OMR software is Audiveris (Java) (https://audiveris.dev.java.net/).

j) Digital moving images and video

Video is the technology of electronically capturing, recording, processing, storing, transmitting and reconstructing a sequence (frames) of still images representing scenes in (full) motion.[43] Digitally born videos can be captured with a cell phone, a digital camera with the necessary applications built into it, and a digital video camera.

Flash video files have a .flv file extension and can be played by VLC (open source, free software media player written by the VideoLAN project) or QuickTime and Windows Media Player (flv-aware players).Some common file formats used for video files are: .3g2 (3GPP2 Multimedia File); .3gp (3GPP Multimedia File), .asx (Microsoft ASF Redirector File), .avi (Audio Video Interleave File), .mp4 (MPEG-4 Video File), .mpg (MPEG Video File) and .rm (Real Media File). A detailed guide on file extensions can be found on the Internet at http://www.fileinfo.com.

k) Animation, interactive (Flash)

Animation is a simulation of movement created by displaying a series of pictures or frames and is often used in multimedia presentations. It is normally vector-based and created with special software which can be viewed by programs freely available from the Web.

The difference between animation and video is that video takes continuous motion and breaks it up into frames; animation starts with frames and puts them together to form the illusion of continuous motion.

The flash format has become a popular method for adding animation and interactivity to web pages and the Adobe Flash Player is available free for electronic devices. The Graphic Interchange Format (GIF) is also often used for animation objects.

l) Games, various genre (.swf)

This term refers to computer generated electronic games and is born-digital content. The ShockWave Flash format (.swf) is traditionally used and can be played in a stand alone Flash Player or be incorporated into a projector or a self-executing Flash movie (.exe extension). Adobe Flash is proprietary software.

The 'Digital Services' website, of the National Library of Australia[44], contains valuable information on storage and management of digital objects.

4. Viewing of file formats

The most often used output device on a computer is the monitor or display unit. A monitor gives the user instant feedback by viewing the text or multi-media. A monitor should regularly be calibrated. Two types of monitors are available; the older cathode ray tube (CRT) technology and the liquid crystal display (LCD) which is not the default option with most computers. The same type of connection to the computer is used for both these technologies. The viewing size, colour depth, bit depth and resolution vary in size for monitors. Monitors also have settings available such as contrast enhancement, texture enhancement and colour correction.

A monitor's *gamma setting* determines the brightness of mid-tones displayed by the monitor and it is advisable to calibrate the computer and attached hardware before working with digital documents. Although the digital creation of a physical object might represent a true facsimile of the original content, the display thereof might differ from monitor to monitor.

5. Preservation of digital objects (see Chapter 7: Preservation)

Migration and emulation are both preservation strategies that are used to ensure the accessibility of digital objects over time.

5.1 Migration

The migration of a digital object involves the transfer of data to a newer system environment, e.g. copying data from magnetic tape to DVD or CD (Digital Versatile/Video Disc/Compact Disk), for preservation purposes.

---

42 Fort Notation software, http://www.forte-notation.eu. (accessed January 15, 2010)
43 Wikipedia, http://en.wikipedia.org/wiki/Video. (accessed 15 January 2010)
44 Digital Services, National Library of Australia, http://www.nla.gov.au/dsp/ and http://www.nla.gov.au/initiatives/digarch.html. (accessed January 15, 2010).

5.2 Emulation

Emulation involves the simulation of the hardware and software used at the creation of digital objects to replicate the functionality of an obsolete system onto a newer platform, e.g. a word processing application no longer available on a specific operating system platform.

Migration and emulation will require detailed information about the format which can be embedded in the document or stored separately as a 'side-car' containing relationships to the master document.

6. File format registries

File format registries contain information about file formats, as well as the hardware platforms and software applications that support them. Two of the most well-known examples are:

6.1 PRONOM
(http://www.nationalarchives.gov.uk/aboutapps/pronom/)

The National Archives of the United Kingdom[45] has developed an online registry called PRONOM[46] which is an on-line information system about data file formats and the supporting software products required to support long-term access to electronic records and other digital objects. PRONOM holds information about file formats and the software products which can process (read, write, identify, etc.) each format. The registry initiative of the UK National Archives joined forces with the Global Digital Format Registry (GDFR)[47] initiative to form a strong single formats registry - the Unified Digital Formats Registry[48].

6.2 XENA (http://xena.sourceforge.net/)

XML Electronic Normalising for Archives (XENA) is an open source software tool which can be used to assist in the long-term preservation of digital objects. XENA aids digital preservation by detecting the file formats of digital objects and converting them into open formats for preservation. This tool was developed by the National Archives of Australia.

Various standards exist for file formats and are freely available from the Internet. It is best to choose formats that are not proprietary dependent and are widely used. No digital format will last forever, but standardisation will improve the chances for preservation. Master objects should be kept in the original format and structure of creation.

7. Digital imaging guidelines

Digital imaging differs from one project to another, the set standard, purpose and method of its use as well as its intended audience are factors that should be taken into consideration for creating digital images. The quality and condition of the original, physical object also impacts on the resolution of capture and the end quality of the digital image.

There is a direct correlation between the production quality of a digitised object and the readiness and flexibility with which that object may be used, reused and migrated across platforms (NISO: Object Principle 1)

7.1 Master Files (High-quality archival image)

Recommended file formats to create high-quality archival images are TIFF or JPEG2000. The master image should be of the highest affordable quality, should not be edited or processed at all. Uncompressed and intensive quality control should be applied when master image files are created, with technical metadata captured as part of the file structure. Master images should be stored in a non-proprietary file format which supports the fidelity and long-term preservation of the image.

7.2 Service Master Files (Working copy of the master image)

A TIFF copy should be made in order to create a lasting working copy of the master image. Photo editing software such as Adobe Photoshop may be used for the necessary image editing. Procedures can include (if necessary): rotation, crop, contrast settings, or colour management and sharpening, depending on the project goals. While working on these files the lossless storage format is preferred.

7.3 Derivative Files (Used for presentation)

Recommended file formats as described in this chapter can differ, but the general file format output for images is JPEG. The derivative files are created from the service master and include the access image and thumbnail. These files are linked to the master image by a unique file name.

8. Metadata

Definition: 'Metadata is structured information associated with an object, for purposes of discovery, description, use, management, and preservation.' (NISO 2007)

A good object should be meaningful and useful outside of its local context and contain associated metadata in the form of descriptive, structural and administrative metadata. The metadata of an object should be self-contained, and include pertinent information about the object, such as technical information for its use, rights management and fixity information. Fixity information contains digital signatures, passwords, etc. The master object should not be bound by passwords or restrictions to open the file format.

Several metadata schemas exist and standards for the schemas are available from the World Wide Web and International Standard Organisations. (see also Chapter 5: Metadata)

45 The National Archives, http://www.nationalarchives.gov.uk. (accessed January 15, 2010).
46 PRONOM, http://www.nationalarchives.gov.uk/PRONOM. (accessed January 15, 2010).
47 Global Digital Format Registry, http://www.gdfr.info/. (accessed January 15, 2010).
48 Unified Digital Formats Registry, http://www.udfr.org/. (accessed January 15, 2010).

9. Identifiers

For storage and Web display of an object, persistent identifiers are used to ensure retrieval and access throughout the lifecycle of the object. An intermediate system needs to 'resolve' the identifier to the correct physical location where the object is stored. The identification system will update a record when the physical location for an item changes in the resolution system.

The software architecture of a digital object provides for a unique identifier i.e. the file name, which contain an identifier that uniquely specifies a single digital object within the parent collection, and the filename extension which identifies the format of storage. A good object will be named with a persistent, (at the minimum) locally unique identifier that can be resolved to the current address of the object during its intended life span.

It is possible to incorporate standard identifiers into a local naming scheme by including, as a prefix, the organisation where the filename was assigned. This is the case with globally unique identifiers which need prefix elements, such as a code representing the organisation, as an addition to the name.

A name resolver is special software that uses a registry to map from the static persistent identifier to the current location of the object. A Persistent Uniform Resource Locator (PURL) does not directly describe the location of the resource to be retrieved, but rather acts as a helper for modules using elements of the object request to sustain information between objects. PURLs are interim measures while Uniform Resource Names (URNs) are constantly maintained.

9.1 Handle System

A 'handle' is a persistent identifier of arbitrary resources. The Handle System is a general purpose distributed information system that provides enough extensible and secure information for the object to be identified on networks, i.e. the Internet. It contains information elements about the creation method of the object, its unique identifier and metadata about the object. The system is managed by the Corporation for National Research Initiatives (CNRI)[49]. Any organisation that wishes to use the Handle System technology, must first register with the CNRI. The CNRI in turn issues a unique prefix identifier schema, to a participating organisation from which item-level identifiers are assigned on application level (e.g. DSpace Handle software). The combination of the institutional Handle prefix, with the unique item-level identifiers, results in the persistent Handle.net identifiers.

9.2 Digital Object Identifier (DOI)

The DOI is a managed system for persistent identification of content on digital networks. It can be used to identify physical, digital or abstract entities. The identifiers (DOI names) resolve to data specified by the registrant, and use an extensible metadata model to associate descriptive and other elements of data with the DOI name.

9.3 Universally Unique Identifier (UUID)

The intention of the Universally Unique Identifier (UUID) URN namespace, also known as GUID (Globally Unique Identifier) is to enable distributed systems to identify information uniquely without significant central coordination (as is the case with the Handle.net and DOI systems). Thus, anyone can create a UUID and use it to identify something with reasonable confidence that the identifier will never be unintentionally used by anyone for anything else.

9.4 File naming

Digital objects need to be organised and named correctly to ensure that they will be identifiable and accessible. Efficient management of electronic records begins with accurate file naming. It is good practice to use file naming that will be recognised in as many different environments as possible. The scheme used in the file naming convention should be documented and the documentation should be accessible.

Good practice is
- to avoid special characters such as: \ / : * ? ' < > [ ] # $ %;
- to use under-scores instead of periods or spaces;
- that each filename should be restricted to 8 characters (it can run up to 25 characters for descriptive information, if necessary);
- to use dates consistently;
- to include a version number, when applicable, for preservation purposes;
- to include enough descriptive information to be recognised if objects are pulled out of their folders;
- not to include spaces in the file name;
- to include only ASCII letters ('a' through 'z'), ASCII digits ('0' through '9'), hyphens, underscores and periods in the file name;
- to be consistent.

Each collection will impose its own restrictions on file names. Processes must in be in place to identify easily which file refers to a specific digital object, and the position it occupies within that object.

The derivative file of a master object must have the same name as the master file, with one exception to indicate it as a derivative file. A derivative file will typically have a different file type with its own unique format extension. It should always be easy to identify files with master-derivative relationships.

The file naming scheme an institution decides to use depends on the limitation of the computer system used for the creation of a digital object and the types of networks with which the institution intends to exchange files. Limitations of any backup or off-site storage system will also influence the file naming convention which should ensure the future retrieval of a digital object.

To allow for the storage of multiple versions of the same document on a computer one should add a number or alphabet letter at the end of the file name to indicate the version.

---

[49] The CNRI Handle System. Available from: http://www.handle.net/. (accessed January 15, 2010).

10. Sustainability

The term 'sustainable format' means the ability to access an electronic record throughout its lifecycle, regardless of the technology used when it was originally created. A sustainable format is one that increases the likelihood of a record being accessible in the future.

It is important to take note of the fact that there are some sustainability factors that apply to all digital formats and which influence the feasibility and cost of preserving content in the face of future change. These factors include, amongst others, disclosure, adoption and transparency considerations.

For a complete listing and explanation of sustainability factors, please refer to the overview of digital formats and sustainability factors that apply to all digital formats provided by the National Digital Information Infrastructure and Preservation Program (NDIIPP) of the United States of America. This information is available in online format at http://www.digitalpreservation.gov/formats/sustain/sustain.shtml.

11. Authentication

Digital objects bear less evidence of authorship and proof of provenance than their physical counterparts. Authenticity features for digital objects consist of documentation about the origin, chain of custody, relationship to other objects and characteristics of an object. The user does not know if or when a digital object has been changed since its creation, or if the alterations changed the fundamental essence of the object, for this reason a means for indicating its authenticity needs to accompany a digital object. The integrity and trustworthiness of a digital object refers to the degree of confidence a user can have when using it.

Digital objects have attributes that can be used to assist authentication. It is important to note that authenticity does not refer to the accuracy of the content or meaning of the object. Usually, the context of the original object can be related to its other versions, as well as to other objects within a collection. The master object should serve as the main source of authenticity.

Fixity of an object can include licenses, methods of use and restrictions to access. Digital watermarking usually indicates the copyright state of a digital object, but can also be used to add proof of authentication to a digital object. However, it is important to note that watermarking tampers with the original object and can sometimes damage the visibility of the content.

Digital objects are defined as a set of sequences of bits. A checksum or hash-sum consists of fixed-sized algorithms which can detect accidental errors that may have been introduced during the handling of a digital object. Checksums allow users to consider the integrity of various versions of an object. The integrity of the data can be checked at any time during the life span of a digital object and compared with the archival object. If the checksums do not match, it means that the data has been altered.

The following figure shows a typical checksum function.

Figure 4.5 a typical checksum function Source: http://en.wikipedia.org/wiki/File:Checksum.svg
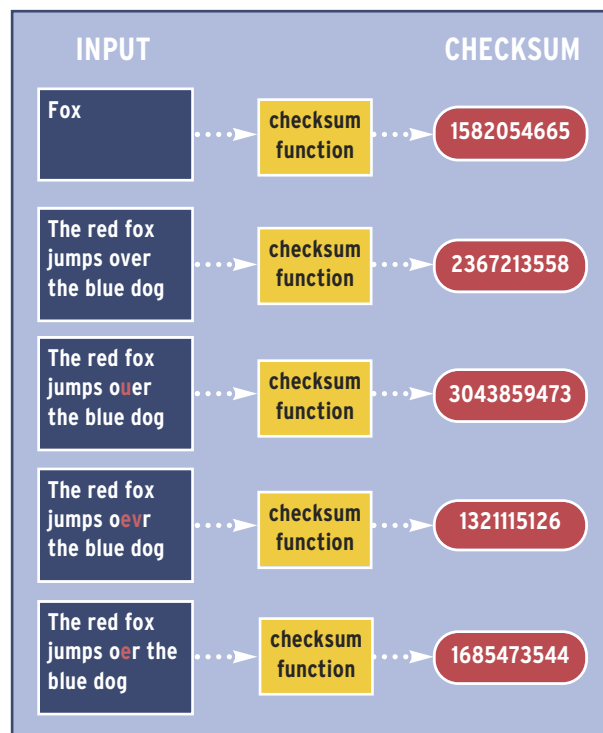
12. Conclusion

This chapter focuses on the advantages and responsibilities involved in the use of reusable objects. Objects can be of great value to a digital collection, provided the formats used to house them are applicable for the intended audience, can be viewed by the majority of users and are named in such a way that they enhance the sustainability of the collection.

References

Hughes, L.M. (2004). *Digitizing Collections, strategic issues for the information manager*. New York: Facet,.

Library of Congress, Digital Preservation (nd). *Sustainability of Digital Formats: Planning for Library of Congress Collections*. Available from: http://www.digitalpreservation.gov/formats/sustain/sustain.shtml (accessed 7 April 2010)

National Archives and Records Administration, U.K. (June 2004). *Technical Guidelines for Digitizing Archival Materials for Electronic Access: Creation of Production Master Files – Raster Images*. Available from: http://www.archives.gov/preservation/technical/guidelines.html (accessed 7 April 2010)

NDIIP.(nd). *Sustainability of Digital Formats: Planning for Library of Congress Collections* Available from: http://www.digitalpreservation.gov/formats/sustain/sustain.shtml. (accessed 07 April 2010)

Townsend, S., Chappell, C., and Struijvé, O. AHDS Guides to Good Practice, *Digitising History: A Guide to Creating Digital Resources from Historical Documents* Available from: http://hds.essex.ac.uk/g2gp/digitising_history/index.asp. (accessed 07 April 2010)

Chapter 5

Metadata

1. Introduction

Every day we consult and use directories, maps, guides, indexes and other lists to assist us in finding physical resources and information. In libraries and art galleries, catalogues perform the same function. Information on product-packaging and medicine containers enable us to make informed decisions about our purchases.

In an increasingly digital environment, information and resources that we seek are stored on computers and servers, on the Internet and the World Wide Web. Information resources created and stored digitally, require discovery, identification and access in much the same way as sourcing information in a physical environment. Assistance in finding digital resources can be provided by creating descriptions of the resources themselves using structured, consistent but concise information associated with those resources. This is known as creating metadata.

This chapter explores the concept of metadata and its role and function in a digital environment. International standards, best practices and local guidelines for selection of appropriate schemes underpin successful exchange and reuse of quality metadata, thereby enabling global networking of information resources.

2. What exactly is metadata?

Metadata is therefore literally 'data about data'. It is not the resource object itself but a description of the object. Metadata is structured information, in digital format, describing specific digital resources stored as computer data, readable by humans but also understood by computers. Metadata made available on the Web, enables resources, whether digital or physical, to be discovered, usually as a result of search engines and Web crawlers such as Google, Yahoo and others, discovering and indexing resources and making them 'visible'. The resource object itself, may or may not (often depending on the rights attached to the object) be made available to users of the Web.

3. Purpose and function of metadata

Metadata is multifunctional but, generally speaking, is created for the discovery, access, identification, management and preservation of digital objects, particularly on the World Wide Web. High quality metadata that is standardised and consistent may be used and shared. It may also be created on an incremental basis, adding detail and 'richness'.

Currently a large proportion of metadata is created manually, but it will increasingly become software-driven in automated processes.

4. Different types for different functions

Differentiation of metadata into types based on their function is helpful, especially when creating metadata is a shared responsibility. Metadata may be created for collection level description and/or item level description.

4.1 Descriptive metadata

Descriptive metadata is information about resource titles, creators, dates, often found on the resource itself. Added information about subjects, places and time periods relating to the intellectual content of the resource is often included. Librarians, cataloguers, personal authors and publishers normally create descriptive metadata.

Therefore, descriptive metadata is used for correct and unambiguous identification, location, access to and discovery of specific resource objects. Correct identification may be enabled, for example, by the use of an ISSN (International Standard Serial Number), an ISBN (International Standard Book Number) or other formal identification. Location and access can be facilitated by the use of a URL (Uniform Resource Locator), URI (Uniform Resource Indicator) to point to the location of the resource that is available on the Web, or to the institution where the physical resource may be housed. Subject headings, keywords, thesauri and taxonomies enable users to discover resources on the same or related subjects. Descriptive metadata is usually made available to users.

4.2 Administrative metadata

Administrative metadata includes information about the creation of digital files, such as file formats, scanning dates, file compression formats, copyright and other rights. Rights metadata may include information about copyright holders, permission to use statements, limitations and restrictions of use, and use disclaimers. Recording of technical information such as the hardware and software used to create a digital object (either converted from physical to digital or born-digital) allows files to be migrated to new formats, facilitating long-term preservation (see the section on Preservation for more information). Some of this metadata may be captured by software as part of the digital file creation, and others, such as rights information, may be manually created. Therefore, administrative metadata enables digital file administration, curation and management, rights monitoring and management, as well as the preservation of resources, over a period of time. Administrative metadata facilitates the management of digital resources and would not necessarily be made public apart

from terms and conditions of use and copyright ownership.

4.3 Structural metadata

Structural metadata captures information about the structure and relationships between digital objects such as the scanned pages of a book, related resources in a collection, and the use of unique and persistent identifiers. Thus structural metadata ensures the integrity and usefulness of resources over a period of time by maintaining the technical relationship between the component parts.

5. Standardisation

In order to standardise the creation of metadata, ensure consistency and enable the sharing and re-use of metadata it is essential to select a metadata standard that is appropriate to the type of resource and the community of users. The application of standards is essential for isolated digital collections to move to interoperable collections and for libraries to move into a larger standards world. Open standards are themselves becoming increasingly interoperable. Metadata standardisation is fundamentally the most important step towards interoperability.

Standards for metadata are created by communities of experts and can be proprietary or open. Proprietary standards are those developed through commercial and market use and for which proprietary software is required together with purchase and license fees. Open standards are those created through shared and collaborative initiatives by experts in a specific community and made freely available. Open standards evolve with community application, use and adaption over a period of time. Adopting open standards promotes collaboration between institutions, facilitates exchange and reuse of metadata among similar communities, reduces fragmentation of efforts and, importantly, reduces the costs associated with creation of metadata.

5.1 Data structure standards

These are essentially structured containers for metadata representing an object or a collection. The choice of elements and the rules of application are clearly defined within the standard and illustrative sample records are available from the websites listed with the standards. The most well known and used data structure standards are:

- MARC (Machine-Readable Cataloguing Format) - MARC 21; MARCXML http://www.loc.gov/marc/
- DCMES (Dublin Core Metadata Element Set) – Dublin Core Simple; Dublin Core Qualified http://dublincore.org/documents/dces/
- MODS (Metadata Object Description Scheme) http://www.loc.gov/standards/mods/

- EAD (Encoded Archival Description) http://www.loc.gov/ead/
- TEI (Text Encoded Initiative) http://www.tei-c.org/index.xml
- LOM (*Learning Object Metadata*) http://www.imsglobal.org/metadata/mdv1p3pd/imsmd_bestv1p3pd.html
- METS (Metadata Encoding and Transmission Standard) http://www.loc.gov/mets/
- MIX (Technical Metadata for Digital Still Images Standard) http://www.loc.gov/standards/mix/
- VRA Core (Visual Resources Association data standard for the cultural heritage community) http://www.vraweb.org/projects/vracore4/
- PREMIS (Preservation Metadata Implementation Strategies) http://www.loc.gov/standards/premis/
- Darwin Core (Biodiversity Information Standard) http://rs.tdwg.org/dwc/index.htm
- EML (Ecological Metadata Language) http://knb.ecoinformatics.org/software/eml/
- SANS 1878 (South African Spatial Metadata Standard)

5.2 Data value standards

Data value standards are the terms or 'values' that are used to describe an object or collection. The relevant value is placed within the element structure in the chosen data structure standard. In the example: <dc:creator>John Smith</dc:creator>, John Smith is the creator and his name is the 'value' of the DC element 'creator ', defined in the data structure standard above. It is preferable to select values from controlled vocabularies, controlled lists and thesauri, where possible. This helps to ensure consistency and accuracy of values, minimise spelling errors and keystrokes, enable interoperability and improve search functionality across collections. The best known controlled vocabularies are:

- LCSH (Library of Congress Subject Headings) http://www.loc.gov/cds/lcsh.html
- MeSH (Medical Subject Headings) http://www.nlm.nih.gov/mesh/
- AAT (Art and Architecture Thesaurus) http://www.getty.edu/research/conducting_research/vocabularies/aat/
- TGN (Getty Thesaurus of Geographic Names) http://www.getty.edu/research/conducting_research/vocabularies/tgn/

5.3 Data content standards

These are the standards which inform the structure of the value, i.e. 'rules' for the way in which you enter values within the data structure container. For instance, the creator, above, may be entered as <dc:creator>John Smith</dc:creator> or <dc:creator>Smith, John</dc:creator> depending on the choice of 'rules'. These rules need to be documented. The best known data content standards are:

- AACR (*Anglo-American Cataloguing Rules*)
- RDA (*Resource Description and Access*) http://www.rdaonline.org/
- RAD (*Rules for Archival Description*) http://www.cdncouncilarchives.ca/archdesrules.html
- FRBR (*Functional Requirements for Bibliographic Records*) http://www.ifla.org/VII/d4/dbc.htm
- FRAD (*Functional Requirements for Authority Data formerly Functional Requirements for Authority Records*) http://www.ifla.org/VII/d4/dbc.htm
- FGDC (*Federal Geographic Data Committee*) http://www.fgdc.gov/metadata/csdgm/

5.4 Schemes and element sets

A metadata scheme or element set is a predefined set of elements for the creation of metadata within and for a specific community, based on adopted community standards. The scheme defines the relationship between the individual metadata elements. The term scheme is used to refer to a text description while schema is usually a diagrammatic representation of the relationship. There is a great variety of schemes available for the description of digital resources, some of which are mentioned in Parargraph 5.1, Data Structure Standards above. Careful comparison and assessment is required to ensure that the most suitable scheme is selected. More than one scheme may be required to reflect objects or collections adequately. For example, DC may be selected for the descriptive metadata, PREMIS for the preservation metadata and METS as the 'container' to link the structure of the object with its metadata and to provide for the exchange of objects between repositories. Considerations should include the following factors:

- Nature of the collection (photographs, newspapers, archival documents, literary works, etc);
- Individual objects and the type of materials;
- Community of users;
- Required granularity of metadata (i.e. collection level or item level);,
- Skills, resources and expertise of metadata creators;
- Purpose (access, preservation, exchange, etc);
- Potential user;
- Technical infrastructure available to the institution;
- Budget.

To ensure interoperability between collections across institutions, it is important to select a scheme that is in use by similar institutions. The websites of schemes generally include practical examples of encoded records for reference purposes.

6. Application profiles, guidelines and best practice

Groups of practitioners sharing a common interest with a willingness to work together in finding solutions to common problems are known as Communities of Practice. Similarly groups of resources that may require similar descriptions and similar vocabularies could use an Application Profile that has been created to ensure metadata consistency for similar needs.

6.1 Application Profiles

Application Profiles interpret scheme elements, specifically for the terminology and requirements of particular types of resources, for example, cultural heritage materials. This supports the consistency of metadata, within and across institutions with the same type of materials, enables more successful sharing of metadata, and improves search results across similar collections.

6.2 Best practice

Best practices are recommendations based on the use of optimal community-defined standards and international standards for selection of schemes, date formats, controlled vocabularies, cataloguing and archival description rules. Best practices provide documentation for decisions regarding formats and standards used. This enables efficient use of tools (technology) to convert, manage and harvest metadata and thereby maintain a standard of quality and promote collaboration.

6.3 Guidelines

Guidelines for metadata creation within an institution help to standardise and improve consistency over a period of time. These should be documented and updated as required. Guidelines should, inter alia, cover the following:

- The definition of elements, in other words, whether elements in a scheme are required, optional, or repeatable;
- Rules for the creation of file names;
- The use of controlled vocabularies;
- Standardised format for dates.
- Guidelines are particularly important if multiple schemes are selected for use within a project or an institution, or if metadata records are to be shared or exchanged through protocols.

7. Encoding

In order for metadata to be understood by computers it is necessary to 'mark-up ' or encode the metadata in such a way that it is machine readable. This is done by encoding in a mark-up language. The structure of mark-up languages relies on tags, opening and closing, in order to label the elements, for example <title> </title> where the actual title, known as the value, is inserted between the tags. A slash ('/') in the second tag indicates the end of the value. Multiple occurrences of an element need to be in repeated tags.
Example: <creator>John Smith</creator>
                            <creator>Paul Brown</creator>
The most well known encoding language is HTML (Hypertext Mark-up Language). HTML is traditionally used for the creation of web pages, allows for hyperlinking and browsing but does not enable complex and intelligent searching. Software editors are available for creating HTML and HTML web pages.

7.1 Extensible Markup language, (XML)
Extensible Markup language (XML) has become the de facto standard for encoding metadata for the World Wide Web. In a way XML resembles HTML and like HTML, makes use of *tags* (words bracketed by '<' and '>') and *attributes* (of the form name equals 'value'). While HTML specifies what each tag and attribute means, and how the text between them

will look in a browser, XML uses the tags only to delimit pieces of data, and leaves the interpretation of the data completely to the application that reads it.

7.2 Characteristics of XML

- **XML is used for structuring data.**
  Structured data includes things like spreadsheets, financial transactions and technical drawings. XML is a set of rules for designing text formats that allow data to be structured. XML is not a programming language, and one does not have to be a programmer to use it or to learn it. XML makes it easy for a computer to generate data, read data, and ensure that the data structure is unambiguous.

- **XML is text but is not read.**
  XML stores data in text format which allows one to look at the data, using a text editor. XML files are text files that one should not have to read, but may be read when the need arises. XML specification forbids applications from trying to second-guess the creator of a broken XML file. If the file is broken, an application stops immediately and reports an error. To ensure that an XML file can be read correctly by a computer, it is required to be well formed, in other words, it needs to comply with a set of pre-defined requirements. To ensure that the XML file complies with pre-defined requirements of a metadata scheme, it has to pass a validation test. The many XML editors that are available to create XML metadata files will perform these tests and report any irregularities.

- **XML is modular.**
  XML allows one to define a new document format by combining and reusing other formats. Since two formats developed independently may have elements or attributes with the same name, care must be taken when these formats are combined (for instance, does '<p>' mean 'paragraph' from this format or 'person' from another?). To eliminate name confusion when combining formats, XML provides a namespace[50] mechanism. XML Scheme is designed to mirror this support for modularity at the level of defining XML document structures, by making it easy to combine two schemes to produce a third which covers a merged document structure.

- **XML enables mapping.**
  Metadata that is created in XML format can be mapped between schemes, and transformed from, for example, MARCXML (MARC 21 XML Scheme) to Dublin Core (DC) and multiple outputs are able to be created in readable formats such as HTML or XHTML for web display and PDF for printing.

---

50 http://www.w3.org/TR/REC-xml-names
51 Paul Miller
52 http://www.loc.gov/z3950/agency/
53 http://library.caltech.edu/openurl/
54 http://www.openarchives.org/OAI/openarchivesprotocol.html
55 http://www.openarchives.org/OAI/openarchivesprotocol.html
56 http://www.w3.org/
57 Carpenter. L.

- **XML is an open standard.**
  It is license-free and platform-independent and is a recommended format for high confidence long term preservation of text.

8. Interoperability

To be interoperable, 'one should actively be ….ensuring that systems, procedures and culture of an organisation are managed … to maximise opportunities for exchange and re-use of information, whether internally or externally'[51]. Interoperability enables isolated digital collections to grow into networked digital libraries supported by the adoption of open international standards and scheme.

8.1 Semantic level

Digital libraries are becoming less about books and more about the ideas and concepts manifested in the books. The consistent use of subject terms (by using thesauri), subject headings (from a controlled vocabulary) and authority files (for the names of places and people), ensure the consistent and unambiguous description of resources. This ensures not only resource discovery but also access to the intellectual content within resources as envisaged by the development of the Semantic Web.

8.2 Technical interoperability

Technical interoperability refers to what is being exchanged (data elements), how to structure them for exchange (schemes) and how to exchange it (protocols). Protocols for the sharing of metadata include Z39.50[52], OpenURL[53] and OAI-PMH[54]

9. The Open Archives Initiative

The OAI (Open Archives Initiative) http://www.openarchives.org develops and promotes interoperability standards that aim to facilitate the efficient dissemination of content on the Internet.

The OAI Protocol for Metadata Harvesting (OAI-PMH)[55], based on a W3C[56] XML Scheme, is used for exposing and harvesting metadata through a defined http (Web) protocol. See Figure 5.1 below[57].

Figure 5.1. The OAI Protocol for Metadata Harvesting
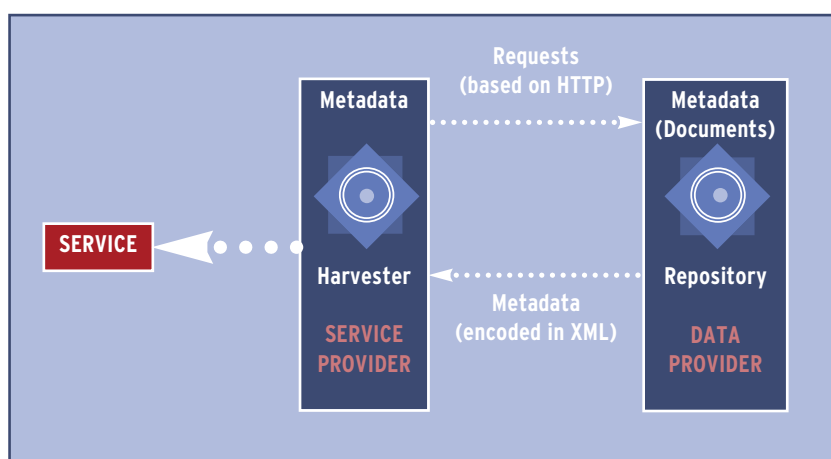
The Protocol works by means of **Data Providers** - open archives, digital libraries, subject repositories institutional repositories, providing free access to their metadata, which may (or may not) include free access to the full texts or other resources. Unqualified Dublin Core (simple DC) is used as the common interoperability metadata. This is either dynamically converted (mapped) from other Scheme to DC, or stored in DC. Mappings between DC, EAD and MARC 21 Schemes are available. All metadata is marked up in XML.

Service Providers run automated harvesting software using the OAI interfaces of the Data Providers to harvest and store metadata. No live search requests are sent to the Data Providers. Value-added services are built on the harvested metadata. Harvesting of records relies on metadata being created according to international standards to ensure compatibility between harvested records from many institutions.

10. Guidance for creating good metadata – NISO's 6 principles

NISO[58], in their document *A framework of guidance for building good digital collections*[59], outline six principles for creating good metadata, which are summarised for ease of reference below. A full narrative explaining the principles are available on the NISO website.

Metadata Principle 1:
Good metadata conforms to community standards in a way that is appropriate to the materials in the collection, users of the collection, and current and potential users of the collection.

Metadata Principle 2:
Good metadata support interoperability.

Metadata Principle 3:
Good metadata uses authority control and content standards to describe objects and collocate related objects.

Metadata Principle 4:
Good metadata includes a clear statement of the conditions and terms of use for the digital object.

Metadata Principle 5:
Good metadata supports the long-term curation and preservation of objects in collections.

Metadata Principle 6:
Good metadata records are objects themselves and therefore should have the   qualities of good objects, including authority, authenticity, archivability, persistence and unique identification.

11. Conclusion

Metadata is increasingly supporting new and innovative services. This underlines the necessity for creating rich and detailed metadata to provide a solid base not only for the long-term management, preservation and delivery of digital resources, but also to facilitate the development of an exciting array of services. The extraction and capturing of metadata will increasingly become more automated with the use of intelligent software. However, this will continue to be supplemented by the skills and knowledge of subject experts, users and researchers.

References

Carpenter. L. OA-Forum Tutorial. (nd) University of Bath, Bath. Available from: http://www.oaforum.org/tutorial/english/page3. htm#section1 (accessed 25 April 2009)

Miller, Paul. (2000) Interoperability: what is it and why should I want it? *Ariadne* Issue 24. Available from: http://www.ariadne.ac.uk/ issue24/interoperability/ (accessed 25 April 2009)

[58] http://www.niso.org/home
[59] http://www.niso.org/publications/rp/framework3.pdf

Chapter 6

Infrastructure

1. Introduction

This chapter is concerned with the digital infrastructure required to host a digital collection, whether born-digital collections or collections digitised from paper, magnetic or other formats. The approach taken in this chapter is to align the composition and make-up of the various components of the digital infrastructure with global best practices and standards, in order to ensure, where practically possible, the future certification of such a digital infrastructure by the criteria and checklists of evolving international benchmarks standards.

1.1 Purpose and Scope

The purpose of this chapter is to provide a set of minimum criteria to which the digital infrastructure of a new or existing digitisation project should adhere.

It is designed to be used as a guideline for creating and maintaining the digital infrastructure for new digital collections, guiding new entrants in the domain of digitisation and digital materials management.

This covers both the system infrastructure required to host digital collections (inclusive of the hardware, software and middleware tier layers) as well as the information content management system designed to ingest, preserve and distribute the content.

1.2 Applicability and Conformance

This chapter is primarily relevant for the more technical oriented persons involved in the management of the digital infrastructure for a digital collections project. However, it does not exclude managerial and other functional-specific staff, like librarians, information specialists, archivists and the like, from studying its contents and become more familiar with the various underlying components of digital collection systems.

In addition, this chapter does not specify a design or digital infrastructure implementation as actual implementations may group or break out functionally different. It is assumed that implementers will use this chapter as a guide while they develop a specific implementation

to provide identified services and content. This chapter does not assume or endorse any specific computing platform, system environment, system design paradigm, system development methodology, database management system, or user interface required for implementation.

2. Standards compliance

A sound digital infrastructure is best expressed by the OAIS reference model (as detailed in the reference framework Blue Book CCSDS 650.0-B-1, adopted as ISO14721:2003)[60]. Such a sound digital infrastructure is in explicit support of the tasks and functions mentioned in the OAIS model, especially in terms of:

- hardware and software,
- formats and storage,
- network and security,
- functions and workflow,
- procedures, protocols, documentation,
- technical and archival skills.

The OAIS reference model is a high-level abstract task model of what has to be accomplished to meet the needs addressed, described in a way that is independent of how it is accomplished. It guides practices, but does not mandate them (for example: it asks that preservation be considered and planned for, but it is does not impose preservation on the implementation as preservation might not be the primary business requirement of a repository).

OAIS is not an architectural benchmark model nor should it be used for certification. It does not guarantee interoperability between systems, but rather largely provides a framework for the standardisation of long-term preservation.

A high-level model is informative of nature and the opposite of a low-level (or normative) model, which focuses on the specifications, orchestration and architecture of system requirements. An example of a low-level model is the e-Framework Service Usage Models (SUMs), that specify more implementation-specific system requirements for developers and implementers.[61]

OAIS as a conceptual model and as a result does not necessarily assist during implementation by focusing purely on the abstract. Because of this it runs the risk of distracting from the real problems that need to be addressed. For this reason the approach taken in this chapter is to

---

[60] Consultative Committee for Space Data Systems, 'Recommendation for space data system standards: reference model for an Open Archival Information System (OAIS)', NASA. Available from: http://public.ccsds.org/publications/archive/650x0b1.pdf. (accessed November 1, 2009).

[61] The e-Framework Partners, e-Framework for education and research. http://www.e-framework.org/. (accessed 1 November, 2009)

supplement the OAIS functional specifications with Trusted Repository attributes, as well as criteria and checklists for Trustworthy Repositories Audit and Certification guidelines[62]. Direct references to the aforementioned guidelines are not made in all cases in the text. In addition, some other real-world practices are referred to where their inclusion is perceived to be beneficial.

3. Digital infrastructure requirements

Any system that supports and administers a digital collection – hereafter more generically referred to as a 'Repository' system – must conform to various requirements that can logically be separated into two different groupings.

The first grouping deals with common services relevant to the successful administration of any type of computerised system. These relate to the network infrastructure of connectivity, hardware, backup, identity and access management services, security considerations, storage and other related functions.

The second grouping is more system-specific in the sense that it deals with the business requirements that such a system tries to resolve, more specifically on the application-level, while it adheres to accepted standards and practices relevant to the environment in which the application functions (see Standards Compliance, Paragraph 2. of this chapter).

3.1 General Services

These common services requirements do not prescribe specific hardware and software, but rather describe good practice for supporting the business application.

3.1.1 Hardware
The Repository requires hardware technologies appropriate to the services it provides to its user community or communities as well as procedures to receive and monitor notifications, and evaluate when hardware technology changes are needed.

The Repository needs to be aware of the types of access services expected by its user community(ies), including, where applicable, the types of media to be delivered (video streaming, graphics, compound data types, documents, text, etc.) and needs to make sure that its hardware capabilities can support these services. For example, it may need to improve its processor capacity, memory usage and networking connectivity (bandwidth) over a period of time to meet the growing access data volumes and expectations.

Hardware can be physical devices housed in a facilities management environment, or a virtualised setup for enabling server consolidation (i.e. one physical device with multiple logical devices operating through a virtualisation tier with the physical device). See Appendix 6.1 at the end of this chapter for a graphical illustration of this concept.

Due consideration should also be given to Sustainable IT criteria for supporting contemporary 'green computing' practices.

3.1.2 Software
The Repository must function on well-supported operating systems and other core infrastructural software. The degree of support required relates to the criticality of the subsystem involved.

Operating systems provide the core services needed to operate and administer the application platform, and provide an interface between application software and the platform.
Services include kernel operations, commands and utilities, security and system management. A choice of proprietary and open source software is available, the preference given to where expertise and functional requirements are most heavily rated.

Core infrastructural software includes the middle tier of a software implementation stack (e.g. LAMP stack = Linux/Apache/MySQL/PHP) like Web, Java servers, or servlet engines as well as persistent identifier services and standards applied for the ingest of content into the Repository. In addition, the choice of application software designed to address the business rules, policies and procedures of a given organisation is most critical. The OAIS functional (OAIS 4.1) and information (OAIS 4.2) models serve as a good departure point in assessing the business functions the application should support.

Application software should also provide support for the long-term maintenance of a bitstream (e.g. to detect bit corruption or loss) as well as the continued accessibility of its contents (e.g. support for linking mechanisms).

3.1.3 Middleware
The middleware services that need due consideration are (i) Identity and Access Management (IAM) and (ii) Security.

IAM services require the usage of, and compatibility with, an enterprise authentication service like LDAP (Lightweight Directory Access Protocol) to allow users of the Repository to authenticate themselves on the system with a set of secure and credible credentials.[63]

Security measures like an enterprise firewall ensuring network perimeter security as well as a server-based firewall are also necessary. Combined with the setting and allocation of appropriate authorisations on file system and database management system level, it will safeguard the Repository against malicious interventions. Significant security updates to the Repository might pertain to software other than core operating systems, such as database applications and Web servers.

3.1.4 Network and Storage
Sufficient network connectivity is needed to allow the Repository to be connected to the Internet in order to allow for the submissions and dissemination of content, the harvesting of descriptive metadata and the resolving of persistent identifier mechanisms.

---

[62] Research Libraries Group, 'Trusted Digital Repositories: Attributes and Responsibilities. An RLG-OCLC Report' OCLC. http://www.oclc.org/programs/ourwork/past/trustedrep/repositories.pdf. (accessed 1 November, 2009)
Center for Research Libraries, 'Trustworthy Repositories Audit and Certification: Criteria and Checklist', CRL. http://www.crl.edu/PDF/trac.pdf. (accessed 1 November, 2009)
[63] Wikipedia, http://en.wikipedia.org/wiki/Lightweight_Directory_Access_Protocol. (accessed 1 November, 2009)

Appropriate space for the storage for file system bitstreams and descriptive metadata in a RDBMS (Relational Database Management System) is a prerequisite for a Repository system. The storage space can be located on the physical hard drive of the server hosting the Repository, a Storage Resource Broker (SRB) device, or it could be centrally located on an enterprise Storage Area Network (SAN). [64]

3.1.5 Backup and Disaster Recovery

The Repository requires adequate hardware and software support for backup functionality sufficient for the repository's services and for the data held, e.g. metadata associated with access controls and repository content.

The Repository content (setup, configuration files, data dictionaries, data sets) can be backed up by using either an enterprise backup management system or a server-based backup system that uses the Operating System commands and scripts. It might be necessary to backup the descriptive and preservation metadata separately from the file system, depending on the overall structure and setup of the Repository.

The Repository must have a suitable written disaster preparedness and recovery plan, including at least one off-site backup of all preserved information together with an off-site copy of the recovery plan. The level of detail in a disaster plan, and the specific risks addressed need to be appropriate to the repository's location and service expectations. This usually takes the form of the Repository being mirrored to a secondary server housed in an alternative facility but at the same time being capable of rendering the same services in real-time mode.

3.2 Application Services

Application services refer to organisational policies that translate to (business) rules that are machine encodable.

3.2.1 Business application foundation: OAIS functional model

The OAIS reference model is a type of archive consisting of an organisation of people and systems that has accepted the responsibility to preserve information for one or more communities. The OAIS contains roles, entities (represented as Information Packages) and functions. Therefore it is clear that the OAIS is not just a technical reference model, but one that includes elements of human behaviour (roles) and data structures (entities) as well.

The foundation of an OAIS digital repository is the Information Package, which includes both a digital object and the necessary associated metadata.

For the purpose of describing a business application designed to meet the functional requirements for managing digital collections, specific focus is now placed on the functional model of OAIS, as graphically illustrated below.

### 3.2.2 Functional requirement: Submission workflows

The Repository content management system must be able to ingest different data types (i.e. receive submissions from the producer/s), depending on the requirements of the digital collections. Data types would typically be simplistic file types (like text documents, graphics, sound clips, etc.) or compound files (like zipped archives). Different packaging standards exist for compound file types, for example:

- METS, a metadata encoding schema expressed in XML format for digital objects ,[65]
- IMS Content Packaging,
- SCORM (Sharable Content Object Reference Model),
- compressed file types (MIME type application/x-zip) for general archiving.

Both SCORM and IMS are standards and specifications for web-based e-learning and packaging e-learning objects.[66]

It is not always feasible to upload vast amounts of content into repository management systems using a web-based HTML interface. Large amounts of data must be submitted into the Repository in batch format, using well-recognised web standards like WebDAV which allow files to be uploaded, edited, deleted and saved in a real-time web environment. A WebDAV-like client which lowers the barriers to deposit content into Repository systems from remote sources, is SWORD (Simple Web-service Offering Repository Deposit), a lightweight protocol for depositing content from one location to another. [67]

Figure 6.1. OAIS Functional Entities
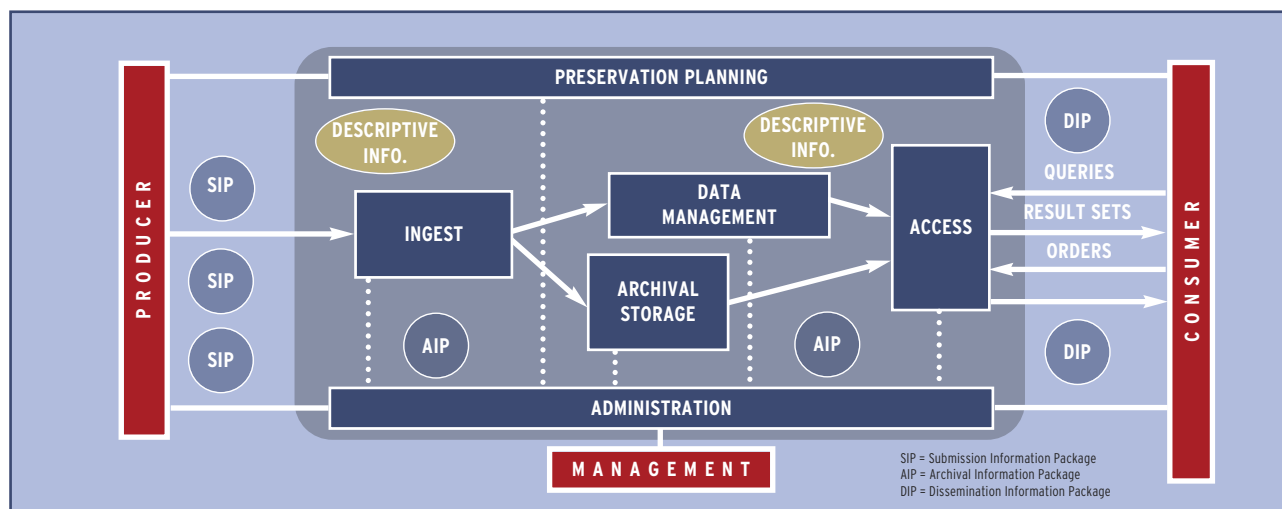
[64] San Diego Supercomputer Center, http://www.sdsc.edu/srb/index.php/Main_Page. (accessed 1 November, 2009)
[65] Library of Congress. http://www.loc.gov/standards/mets/. (accessed 1 November, 2009)
[66] IMS Global Learning Consortium, http://www.imsglobal.org/content/packaging. (accessed 1 November, 2009) SCORM, http://www.scorm.com/scorm-explained. (accessed 1 November, 2009)
[67] JISC and UKOLN, http://www.swordapp.org/. (accessed 1 November, 2009)

Using these compound data types will require a number of pre-ingest activities to prepare the data types according to these predefined standards before submitting them to the Repository. These compound (or simplistic) data types are submitted as Submission Information Packages (SIPs) into the Repository (OAIS 4.1.1.2). It is important that the Repository system should be able to recognise, decode and parse these compound data-type formats. On submission, the SIP is enhanced as necessary and encapsulated as an Archival Information Package (AIP), including the data object and all its associated metadata. Creation of the AIP is the foundation of the long-term preservation function (see paragraph 4 below).

3.2.3 Functional requirement: Archival storage and Data Management

The submission workflow referred to in paragraph 5.2.2 involves, amongst others, the transformation of the object as it was submitted, along with its associated metadata, into a bitstream that can be stored on suitable hardware in the Repository (as well as assigning it a unique identifier). See Appendix 6.1 for a graphic illustration how this can be done on a virtualisation platform. For this reason Archival Storage is a necessary service for the effective storage and retrieval of AIPs, which involve, amongst others, managing the storage hierarchy (i.e. containers and asset stores) and refreshing storage media (i.e. media replacement). It will also include routine integrity checking and basic storage and backup of data.

From a digital infrastructure perspective, Data Management is where the descriptive metadata and system information is stored, more likely in a database. This function is also responsible for maintaining the database, performing queries and generating reports. For repositories this might be an open-source database such as PostgreSQL or MySQL and a series of scripts and configuration files.

3.2.4 Functional requirement: Preservation functions

Long-term digital preservation is a specialised field on its own. It is not the intention of this chapter to expand on this topic (rather see Chapter 7: Preservation). However, in terms of setting up a digital infrastructure to deal successfully with digital preservation concerns, it should be noted that appropriate technologies are required to deal with concerns such as (i) the replication of digital objects and accompanying metadata, and (ii) the format migration of digital objects. Not all Repository management systems deal with these issues very effectively on their own. A persistent identifier system (such as the CNRI Handle System) is required as an add-on to allow for the continued accessibility of the Repository's digital objects. [68]

3.2.5 Functional requirement: Access and Dissemination services

When a user requests access to an object, a Dissemination Information Package (DIP) is provided, which typically contains a copy of the digital object as well as the necessary metadata and support systems to retrieve and use it. In addition to the preparation of the DIP, dissemination requires mechanisms for both verifying the integrity of the information in the DIP and for ensuring that users have permission to access the material. Digital materials cannot be considered preserved without meaningful access, especially by user communities that might require different levels of access to particular materials at different times.

4. Interoperability

A good repository management system conforms to the concept of goodness, which is built upon the principles of 'interoperability', 'reusability' and 'to build services upon' to mention just a few.

Vocabulary and definition:
interoperable *adj.*
> able to operate in conjunction
> interoperability *n.*

(Concise Oxford Dictionary, 9th Edition)

interoperability
> [...] is the ability of a system or a product to work with other systems or products without special effort on the part of the customer. Interoperability becomes a quality of increasing importance for information technology products as the concept that 'The network is the computer' becomes a reality. For this reason, the term is widely used in product marketing descriptions.
> (whatis.com)

The rationale behind any project related to managing digital collections is to make digital materials available to user communities as seamlessly as possible. For this purpose various network services have been developed to enable the sharing of metadata and digital objects between repositories, and between users and repositories.

4.1 Data exchange

The Repository must support the exchange of metadata and digital objects by using appropriate Internet protocols like Z39.50[69], OAI-PMH[70], OpenURL and XML. These protocols can interface between client machines and the Repository over the HTTP (Hypertext Transfer Protocol) which is used for web-based communication over the Internet. The OAI-PMH data provider has been designed specifically for this purpose and acts like a RESTful web service for inter-server communication. In addition, the Repository will require a set of batch Import and Export scripts to allow for data exchange between disparate repositories using a common interface protocol, like METS (metadata wrapped in XML-format).

Digital objects which are presented in a repository as bitstreams must be shareable (governed by access control restrictions) by referencing unique identifiers assigned to such bitstreams. Such data elements must be open to be crawled not only by OAI harvesters, but also by Search Engine spiders and web crawlers (like GoogleBot, Heritrix, etc.).

4.2 Systems integration

A Repository system needs to become part of the larger enterprise setup for any given organisation. This requires repository systems to

---

[68] Corporation for National Research Initiatives (CNRI), http://www.handle.net. (accessed November 1, 2009)

[69] International Standard Maintenance Agency, Library of Congress. http://www.loc.gov/z3950/agency/. (accessed 1 November, 2009)

[70] Open Archives Initiative, http://www.openarchives.org/. (accessed 1 November, 2009)

expose a set of business logic public APIs (Application Programming Interface) to allow implementers to develop code using those APIs to interface directly with the repository's content management code classes. In addition, the Repository has to be compatible with standard authentication protocols like LDAP and Shibboleth to integrate fully with enterprise management systems.

4.3 Reusability

A good repository system that caters for more than just the basic set of services (like preservation, access, etc.) will allow its set of services and digital objects to be reused in such a way that value is added to existing functionality in the process. This calls for a set of exposed web services and mashup functions to allow external agents to interface with the repository and consume its assets without causing any security compromise in a Web 2.0-type fashion.

5. Conclusion

Everyone involved in the hosting of a digital collection should have some knowledge of the technical aspects necessary. International criteria and benchmarking of standards help to provide guidelines for collections to be regarded as 'good'. Various computing platforms, system environments, designs and paradigms, provide the infrastructure that is required for the maintenance of the collections.

References

Center for Research Libraries (2007). *Trustworthy Repositories Audit and Certification: Criteria and Checklist*, CRL. Available from: http://www.crl.edu/sites/default/files/attachments/pages/trac_0.pdf. (accessed 1 November, 2009)

Consultative Committee for Space Data Systems, (2002). *Recommendation for space data system standards: reference model for an Open Archival Information System (OAIS)*, NASA. Available from: http://public.ccsds.org/publications/archive/650x0b1.pdf. (accessed 1 November, 2009)

NISO Framework Working Group (2007). *A framework of guidance for building good digital collections*, NISO. Available from: http://www.niso.org/publications/rp/framework3.pdf. (accessed 1 November, 2009)

Research Libraries Group, (2002) *Trusted Digital Repositories: Attributes and Responsibilities. An RLG-OCLC Report*, OCLC. Available from: http://www.oclc.org/programs/ourwork/past/trustedrep/repositories.pdf. (accessed November 1, 2009)

Appendix 6.1

Virtualisation technology: server consolidation and storage virtualisation
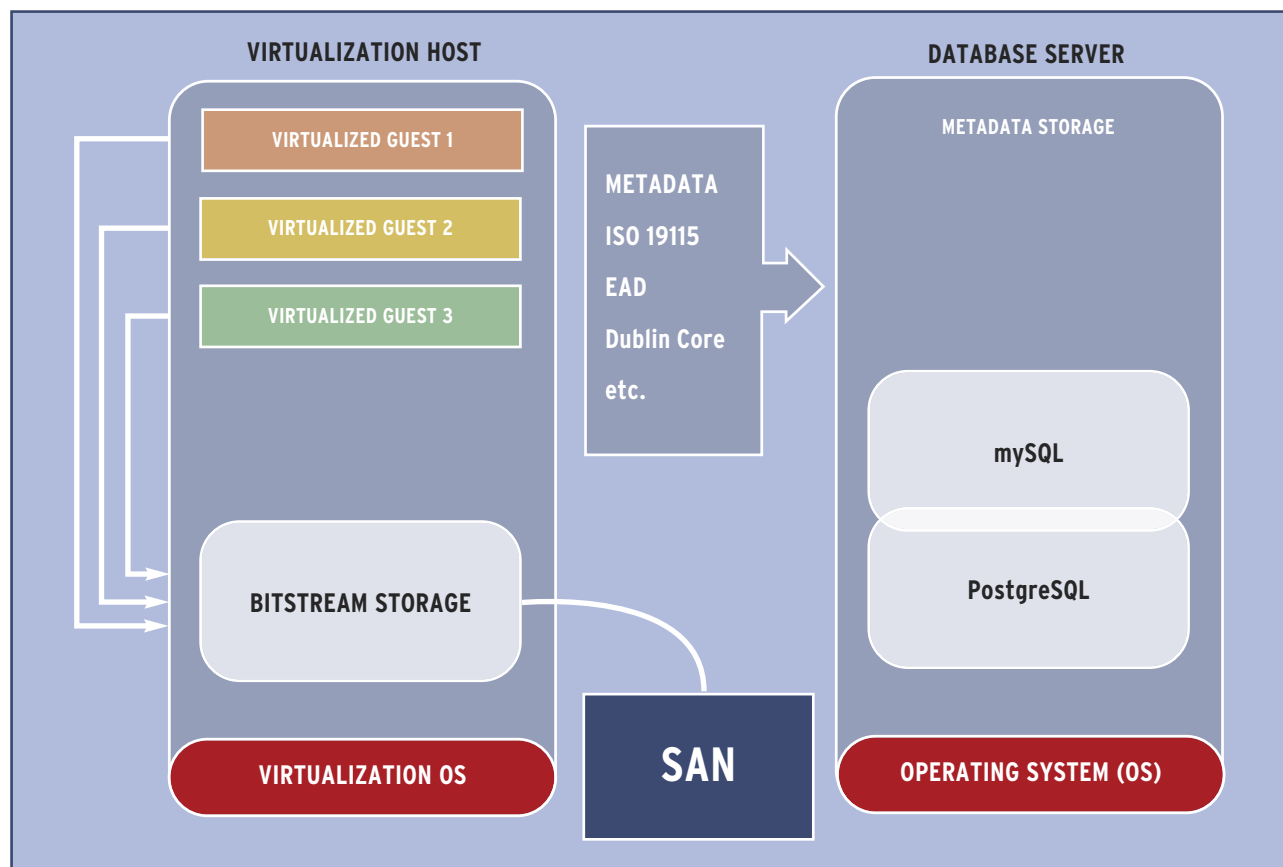
Chapter 7

Digital preservation

1. Introduction

Digital preservation encompasses all of the activities that are undertaken by a digital curator to ensure that the digital content for which the digital curator is responsible is maintained in usable formats across generations of technology and can be made available in meaningful ways to current and future users. An effective digital preservation programme provides an organisation with more than system backup capability and disaster recovery procedures for digital content. A digital preservation programme is designed and implemented to meet the needs of an organisation, from the moment the organisation assumes responsibility for the digital content and for as long as access to the content is required. Introducing good digital preservation practice extends the life of digital content to maximise an organisation's investment in creating digital image collections, even if the organisation might not be committed to providing permanent access to the images. This chapter discusses the components of a digital preservation programme using the attributes of a trusted digital repository as a framework: *administrative responsibility*, *organisational viability*, *financial sustainability*, *technological and procedural suitability*, *system security*, *procedural accountability*, and *OAIS compliance*; and suggests practical means for an organisation to address each attribute.

2. Administrative responsibility

The *administrative responsibility* attribute refers to an explicit commitment by an organisation to preserve specified digital content for which it has accepted responsibility. A mission statement that explicitly refers to or implicitly includes long-term management of digital content is a key indicator that an organisation is meeting this requirement (see Trustworthy Repository Audit and Certification (TRAC) requirements pertaining to governance). For example, the mission statement for the Inter-university Consortium for Political and Social Research (ICPSR) reads:

"ICPSR provides leadership and training in data access, curation, and methods of analysis for a diverse and expanding social science research community.[71]

It has become common practice for an organisation to provide its mission statement on its website.

In establishing its digital preservation programme, an organisation should ensure that the three core aspects of the programme – organisational infrastructure, technological infrastructure, and resources framework – are well-developed and balanced. The organisational infrastructure is best expressed by the Trusted Digital Repository (TDR) document. It is best reflected in the organisation's mission statement and by the development and implementation of required policies as well as the definition of preservation strategies that will be part of the ongoing preservation planning for the organisation's digital collections. The technological infrastructure encompasses the whole of the technical environment, including the hardware and software needed to manage the collections; the file formats and storage media that need to be supported; appropriate levels of network security and controls to ensure that digital content is protected and well-managed; relevant workflows to address the digital preservation functions (as defined in OAIS); documented procedures and protocols for digital preservation activities; and the necessary technical and archival skills to develop and maintain the programme. The resources framework designates the funding for staff, technology, training, and other resources required to sustain the digital preservation programme.

Figure 7.1. Digital preservation three-legged stool.
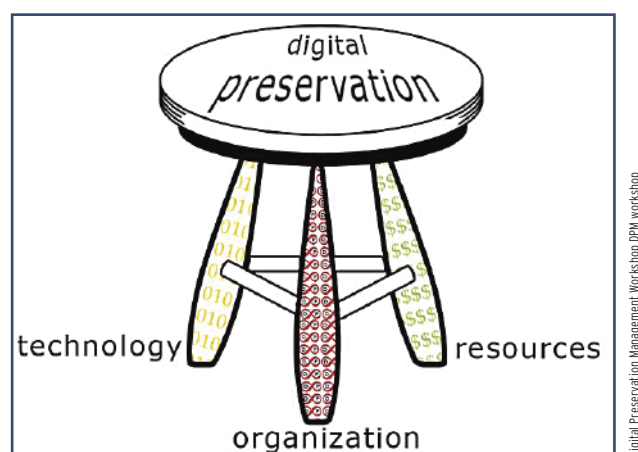
The concept of a three-legged stool, as illustrated in Figure 7.1, conveys the need for balance in the three components of a sustainable digital preservation programme: organisational (what is the scope of the digital preservation programme?), technological (how will the objectives of the programme be met?), and resources (how much will the programme cost the organisation to maintain?).[72]

---

[71] The ICPSR Mission Statement is available at: http://www.icpsr.umich.edu/ICPSR/org/mission.html.
[72] Anne R. Kenney and Nancy Y. McGovern, "Digital Preservation Management: Short-Term Solutions for Long-term Problems," 2003. http://www.icpsr.umich.edu/dpm/

3. Organisational viability

The *organisational viability* attribute addresses the wherewithal of an organisation to preserve specified digital content. Indicators of an organisation addressing this attribute include the existence and availability of policies, procedures, staffing plans, job descriptions, infrastructure plans, and success metrics for the digital preservation programme (see TRAC requirements for organisational structure and staffing and procedural accountability). An organisation engaged in digital preservation should have the appropriate legal status to establish and maintain the programme.

The most tangible evidence that an organisation can provide for this attribute is a comprehensive set of policies that address the full life cycle of digital content management. There are several benefits accrued by an organisation through the process of developing policies. Well-formed policies explicitly define the institutional commitment to the programme; ensure that members of the digital preservation team understand and agree about core aspects of the programme; demonstrate compliance with community standards and practice; manage the expectations of stakeholders, including producers and users of digital content, by being clear about the scope and intent of the programme; identify the issues and challenges the organisation faces in developing the programme; raises awareness about the programme; and identifies the roles and responsibilities inherent in the programme.

Every organisation that is responsible for digital collections should have, at minimum, a high-level digital preservation policy. However, not every organisation should have to invent the components, contents, and structure of such a policy. Appendix 7.1 provides a model for developing a digital preservation policy framework, an organisation's highest-level policy document for digital preservation that de-
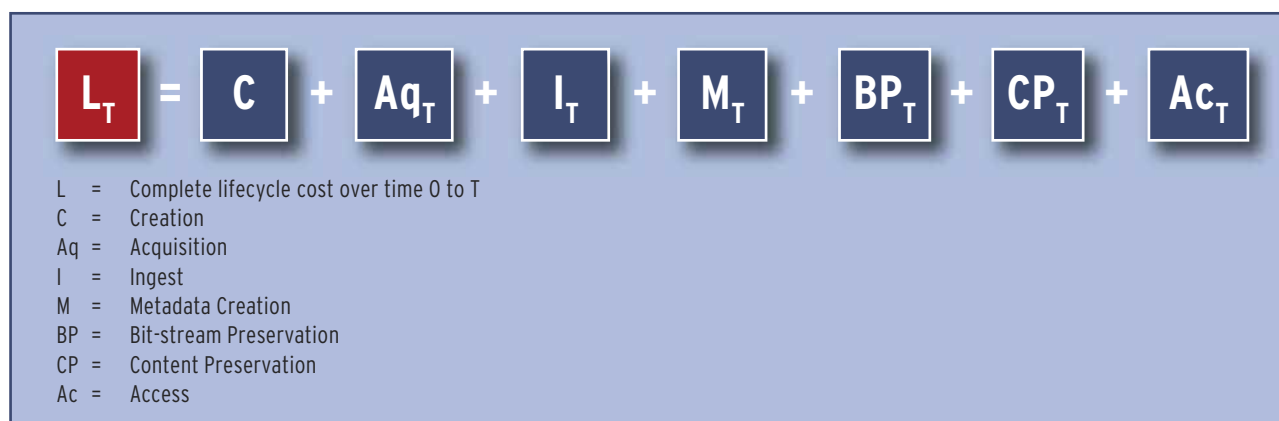
fines the scope and objectives of a digital preservation programme. The model document refers to examples that may be useful in developing a policy that is suited to an organisation's needs for managing digital collections.

4. Financial sustainability

The *financial sustainability* attribute requires an organisation to designate resources for the preservation of specified digital content (see TRAC requirements on financial sustainability for digital repositories). To establish a digital preservation programme, an organisation should anticipate start-up funding, ongoing programme funding, and contingency funding to allow for changes in the scope, content, or technology associated with the programme requiring additional resources. One-time external funding sources may cover start-up and contingency costs, but, typically, organisations need to identify and commit to ongoing funding for the programme.[73] TRAC requires that an organisation be able to demonstrate sustainable, designated funding for digital preservation. Due to annual budget adjustments and other financial shifts that an organisation may face, it may not be possible to provide evidence that funding will be sustained. However, an organisation needs to be able to demonstrate due diligence in funding the programme.

The resources leg of the three-legged stool for digital preservation has recently been the recipient of more intensive community attention. The LIFE project in the United Kingdom has developed a formula for determining the costs of digital preservation, as shown in Figure 7.2.

Figure 7.2. Cost formula for digital preservation developed by the LIFE Project.[74]

$$L_T = C + Aq_T + I_T + M_T + BP_T + CP_T + Ac_T$$

L	=	Complete lifecycle cost over time O to T
C	=	Creation
Aq	=	Acquisition
I	=	Ingest
M	=	Metadata Creation
BP	=	Bit-stream Preservation
CP	=	Content Preservation
Ac	=	Access

LIFE Project

---

[73]   Anne R. Kenney, "Identifying Requisite Resources," session in the Digital Preservation Management workshop series, 2003-2007.

[74]   LIFE Team, LIFE Project, London, UK: British Library and University College London, 2005-2008. http://www.life.ac.uk/2/documentation.shtml.

This formula was initially developed to address the costs of e-journal archiving, but may be applied to other kinds of digital content. It specifies the categories of costs to consider in planning for digital preservation. The digital community is also awaiting the results of the US National Science Foundation's Blue Ribbon Panel on the economic sustainability of digital data, the final report was due in December 2009.[75] These are examples of community developments that should be considered regarding digital preservation costs.

Explanations an organisation might use to account for the costs associated with maintaining a digital preservation programme include:
- Meeting specific obligations (e.g. the programme ensures ongoing access to digital collections that have been created or acquired);
- Maintaining status in relation to peer organisations (e.g. keeping pace with other organisations of the same size and scale that have established a digital preservation programme);
- Avoiding embarrassment (e.g. the outcome that will result if an organisation fails to provide ongoing access to a high-profile digital collection);
- Maximising investments in digital content (e.g. the programme will ensure that the investment in creating digitising collections will be maximised by extending the use of digital content for as long as it is needed or wanted).

5. Technological and procedural suitability

The *Technological and Procedural Suitability* attribute ensures that an organisation's approach is appropriate to its requirements and objectives (TRAC requirements for digital object management). In the past, technology has tended to drive digital preservation developments. Various aspects of technology are considered in other chapters of this Framework (e.g., objects, metadata and digital infrastructure). This discussion considers the sound investment in and management of technology for digital preservation as an organisational priority and requirement. In practice, it is the organisation's requirements that should drive the selection of technology. The first and most important milestone for a digital preservation programme to achieve is the status of well-managed collections, i.e. clear and documented rights for the organisation to have and manage the digital collections over time, digital content stored in well-known and preservable formats, minimal metadata associated with each digital object, as well as controlled and secure access to digital content stored online and offline. If the organisation pursues the implementation of a digital repository to store and manage digital content, that process is easier if the digital content is in a well-managed state.

The technology leg determines the optimal solution for an institution that is based on requirements, resources, priorities, timeframes, and other factors. Suitable technology includes tools, software and workflows that support preservation metadata, digital object packaging, repository software, digital preservation strategies, and other essential aspects of digital preservation. In selecting relevant technology developments to adopt or adapt for an organisation's digital preservation programme, the following factors should be considered. The

development should be written in a well-documented and widely-used language, be usable on a wide variety of platforms, be modular in design, provide support for batch processing of digital objects, be designed for flexible use and integration, be available as open source as often as possible, and be developed by a credible organisation.

6. System security

The *system security* attribute requires that the technical environment used for the preservation of digital content should be effectively adapted and implemented to meet preservation requirements (TRAC requirements for technologies, technical infrastructure, and security). It is essential for a trusted digital repository to be able to demonstrate that its digital content is continuously secure and not vulnerable to technological and other threats. Care needs to be taken by an organisation to meet the specific requirements of managing digital content *over a period of time*. Attention is often focused on immediate, rather than long-term risks, but there are domains of practice to draw upon in addressing these requirements, e.g., information security as well as network security. For example, passing an ISO 27001 Information Security audit satisfies the requirements of the technology section of TRAC.[76]

7. Procedural accountability

The *procedural accountability* attribute enables an organisation to effectively engage in internal self-assessment and external audit and certification (see all TRAC requirements). A digital preservation programme needs to be able to make decisions about how it will preserve its content, document those decisions, document the implementation of those decisions, assess the effectiveness of those decisions, and provide evidence that course corrections are made by the programme based on the assessment. This is referred to as evidenced-based practice. It is not sufficient for an organisation to assert that it is engaging in good digital preservation practice. It must be able to demonstrate to its stakeholders (and ideally also to the community) that it is doing so. Written, approved, and implemented policies and procedures are one of the best and most essential indicators that an organisation is addressing this attribute.

The *Preserving Digital Information* report of 1996 and the OAIS Reference Model both include a call for a certification process for digital archives to demonstrate the effectiveness of the implementation of an OAIS system for preserving digital content. In January 2007, the certification of digital archives became the focus of an international working group to develop an ISO standard via the ISO TC20/SC13 technical committee.[77] The working group used the *Trustworthy Repositories Audit & Certification (TRAC): Criteria and Checklist* that was published in 2007 as a starting point for its work.[78] The work on the certification standard is also informed by the Digital Repository Audit Method Based on Risk Assessment (DRAMBORA) toolkit developed by the Digital Curation Centre and Digital Preservation Europe (DPE), and the work of the nestor project in Germany.[79] Self-assessment, audit and certification efforts define measurable norms for digital preser-

---

[75] See the NSF Blue Ribbon Task Force on Sustainable Digital Preservation and Access, Interim Report of the Blue Ribbon Task Force on Sustainable Digital Preservation and Access, 2008. http://brtf.sdsc.edu/biblio/BRTF_Interim_Report.pdf

[76] See for example, An Introduction to ISO 27001, ISO 27002....ISO 27008, http://www.27000.org/.

[77] Watch for the revision of the TRAC requirements by the ISO working group for public review at: http://wiki.digitalrepositoryauditandcertification.org/bin/view.

[78] The TRAC document is available at: http://www.crl.edu/PDF/trac.pdf.

vation. The self-assessment process produces a development plan for an organisation's digital preservation programme that might require the formalisation of policies, the specification of roles and responsibilities, succession planning for the long-term care of digital collections, a budget review, a metadata inventory, an examination of the transfer of preservation rights to the repository, and technical enhancements to meet requirements.

8. OAIS compliance

The *OAIS Compliance* attribute incorporates OAIS, approved as ISO 14721:2003, into the design and implementation of digital archives (see TRAC requirements for digital object management and technology infrastructure). An organisation that commits to preserving digital content is expected to demonstrate OAIS compliance, which may range from minimally addressing the standard to actively embracing it. The OAIS Reference Model provided the first explicit and the most comprehensive definition and explanation of the activities of digital preservation. OAIS identifies six higher-level activities performed by an archive, each of which consists of individual functions for performing that activity. These include:

- *Ingest*, describing the functional components needed for the secure acceptance and quality control of submissions to an archive;
- *Archival Storage*, explaining the functional components that ensure the secure storage, management, and retrieval of the content of an archive;
- *Data Management*, delineating the functional components for the comprehensive accumulation and provision of administrative data about the operation of and documentation about the content of an archive;
- *Administration*, describing the functional components needed to develop, maintain, and apply the policies and procedures that are used to operate and coordinate the functions of the archive;
- *Preservation Planning*, detailing the functional components needed to develop and recommend standards, policies, procedures, and mechanisms for preserving the content of an archive;
- *Access*, including the functional components to find and deliver content in an archive to authorised users (whether the digital content is for internal or external use).

The OAIS Reference Model is increasingly used to refer to digital preservation activities. The terminology it defines enables discussions that cross the boundaries of the professions, domains, and sectors in which digital content is managed.

9. Implementing a digital preservation programme

In developing its digital preservation programme, an organisation should find the seven TDR attributes to be a useful and informative guide. The attributes have inter-dependencies that inform the implementation of a programme, as illustrated in Figure 7.3.

Figure 7.3. A model of a trusted digital repository.[80]

Digital Preservation Management Workshop DPM workshop
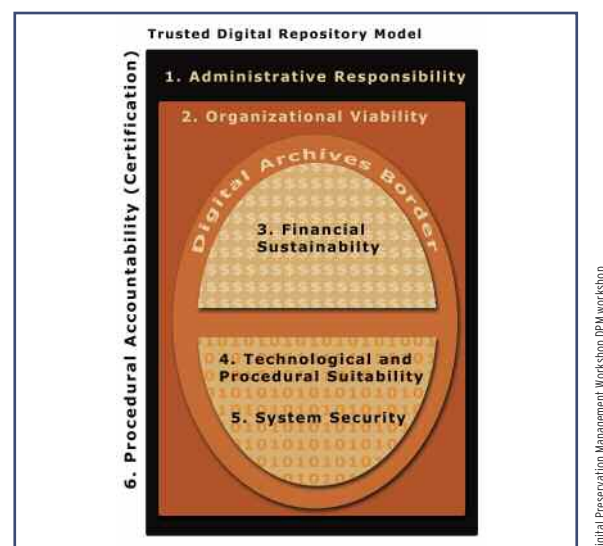
A programme will be difficult or impossible to sustain without an institutional commitment to digital preservation. For this reason the administrative responsibility provides an encompassing wrapper for the programme.

Organisational viability is the next most essential attribute for maintaining a successful programme and it is the most comprehensive of the TDR attributes in scope and impact. Therefore, it is illustrated as the next layer of the digital preservation programme in Figure 7.3. Each instance of a repository (i.e., an individual implementation of repository software to manage specific digital content) must specify the resources (financial sustainability) and the appropriate platform (technological and procedural accountability, and system security); the latter is addressed by OAIS. The next layer represented in the diagram above illustrates digital archives as a border. It is important because an organisation may manage more than one digital repository system or may partner with other organisations to manage a joint repository system – in either case, the outside layers (administrative responsibility and organisational viability) need to be in place, but not repeated and the inside layers need to be specified for each repository system. A trusted digital repository must demonstrate its effectiveness to its stakeholders through self-assessment and audit; thus the procedural accountability layer provides an outer membrane between the repository and the community.

There is no on/off switch for a digital preservation programme, so each organisation must advance through stages of development to achieve a fully implemented programme, for example:[81]
Stage 1. Acknowledge: understand that digital preservation is a local concern.
Stage 2. Act: establish one or more digital preservation projects to address specific issues.
Stage 3. Consolidate: segue or move smoothly from projects to programmes

---

79 Digital Repository Audit Method Based on Risk Assessment (DRAMBORA)', Digital Curation Centre (DCC) and DigitalPreservationEurope (DPE), http://www.repositoryaudit.eu/. nestor Working Group Trusted Repositories - Certification, Catalogue of Criteria for Trusted Digital Repositories, studies 8, Version 1, http://www.langzeitarchivierung.de/index.php?newlang=eng.
80 This model was developed by Nancy Y. McGovern with Anne R. Kenney for the Digital Preservation Management workshop series, 2003, based on analysis of the attributes of a trusted digital repository. The digital archives border was added by the designers of the model.
81 Anne R. Kenney, and Nancy Y. McGovern, 'The Five Organizational Stages of Digital Preservation'. In *Digital Libraries: A Vision for the 21st Century: A Festschrift in Honor of Wendy Lougee on the Occasion of Her Departure from the University of Michigan*, 122-53, 2003.

Stage 4. Institutionalise: incorporate the larger environment and rationalise programmes

Stage 5. Externalise: embrace inter-institutional collaboration and dependency.

These stages provide a means for the organisation to identify incremental steps for developing an organisation's digital preservation programme, provide a way of communicating with peers and others in the community about digital preservation development, and enable measuring progress towards programmatic digital preservation goals.

There are several steps an organisation may take to get started on the implementation of a digital preservation programme. These include the following:

- Bring together a group to develop a digital preservation policy;
- Make an inventory of digital content for which the programme is responsible and work towards achieving well-managed collection status for all of the identified digital content;
- Track relevant community developments and apply outcomes internally;
- Review the institutional mission statement to incorporate digital preservation;
- Conduct a campaign to raise awareness about digital preservation challenges including presentations, discussions, summaries of activities for users, funders, and others;
- Establish a digital preservation task force which includes key stakeholders within the institution;
- Use the TRAC checklist to review the preservation programme to determine what is already in place and what policies and procedures need to be developed.

10. Conclusion

For digital image collections, there are many models and examples from existing programmes to reference in developing an organisation's digital preservation programme. The programme should reflect and conform to the prevailing standards and practice of the digital preservation community to the extent possible. The organisation should develop a prioritised set of objectives for the short-term and long-term development of the programme. The managers of the programme should regularly share updates on progress, identify ongoing challenges the programme faces, and highlight successes in delivering digital collections to users.

References

Beagrie, N. and Jones M. (2007). *Preservation Management of Digital Materials - the Handbook*. Online Edition. York: Digital Preservation Coalition. Available from: http://www.dpconline.org/graphics/handbook/. (accessed 3 June 2010)

Consultative Committee for Space Data Systems (CCSDS). (2002). *Recommendation for Space Data System Standards: Reference Model for an Open Archival Information System (OAIS)*. Blue Book CCSDS 650.0-B-1, no. 1. Available from: http://public.ccsds.org/publications/archive/650x0b1.pdf. (accessed 3 June 2010)

Digital Curation Centre (DCC), and DigitalPreservationEurope (DPE). (2008). *Digital Repository Audit Method Based on Risk Assessment (DRAMBORA)*. Available from: http://www.repositoryaudit.eu/. (accessed 3 June 2010)

Inter-university Consortium for Political and Social Research (ICPSR). (2007). *ICPSR Digital Preservation Policy Framework*. Ann Arbor, Michigan: ICPSR. Available from: http://www.icpsr.umich.edu/DP/policies/dpp-framework.html (accessed 3 June 2010)

Joint Information Systems Committee. (2008). *Digital Preservation Policies Study*. London: JISC. Available from: http://www.jisc.ac.uk/publications/publications/jiscpolicyfinalreport.aspx (accessed 3 June 2010)

Kenney, A. R., and McGovern, N. Y. (2003). The Five Organizational Stages of Digital Preservation. Edited by Hodges, P., Sandler, M., Bonn M. and Price Wilkin, J. *In Digital Libraries: A Vision for the 21st Century: A Festschrift in Honor of Wendy Lougee on the Occasion of Her Departure from the University of Michigan*. Ann Arbor, MI: Scholarly Publishing Office, University of Michigan Library. (122-53)

LIFE Team. (2005-2010). LIFE Project, London, UK: British Library and University College London. Available from: http://www.life.ac.uk/2/documentation.shtml. (accessed 3 June 2010)

McGovern, N. Y. (2007) *A Digital Decade: Where Have We Been and Where Are We Going in Digital Preservation?*. RLG DigiNews 7 (1). Available from: http://worldcat.org/arcviewer/1/OCC/2007/08/08/0000070519/viewer/file137.html#article3 (accessed 3 June 2010)

National Science Foundation Blue Ribbon Task Force on Sustainable Digital Preservation and Access. (2008). Interim Report of the Blue Ribbon Task Force on Sustainable Digital Preservation and Access. Available from: http://brtf.sdsc.edu/biblio/BRTF_Interim_Report.pdf (accessed 3 June 2010)

National Library of Australia and the Digital Preservation Coalition. (1995-on) *Preserving Access to Digital Information (PADI)*: National Library of Australia. Available from: http://www.nla.gov.au/padi/. (accessed 3 June 2010)

Research Library Group (RLG)-National Archives and Records Administration (NARA) Task Force on Digital Repository Certification. (2007). *Trustworthy Repositories Audit & Certification: Criteria and Checklist Ver. 1.0*. Chicago, IL: Center for Research Libraries (CRL). Available from: http://catalog.crl.edu/record=b2212602~S1 (accessed 3 June 2010)

Research Library Group (RLG) and Online Computer Library Center (OCLC). (2002) *Trusted Digital Repositories: Attributes and Responsibilities* Mountain View, CA: Research Library Group (RLG). Available from: http://www.oclc.org/programs/ourwork/past/trustedrep/repositories.pdf. (accessed 3 June 2010)

Appendix 7.1.
Model Digital Preservation Policy Framework

Version 2.0 Digital Preservation Policy Framework: Outline
Prepared by Nancy Y. McGovern, Digital Preservation Officer, ICPSR
January 2007 (last revised October 2008)

Overview

This document provides an outline for constructing the digital preservation policy framework for ICPSR and offers a step towards identifying core components of a digital preservation policy framework to encourage a community standard for digital preservation policy documents. The outline was developed to produce a digital preservation policy framework that:

- Addresses the seven attributes of a Trusted Digital Repository
- Presents the high-level perspective of an organisation's digital preservation program
- Reflects current not future capabilities of the digital preservation program
- Provides links to documents containing more detailed and frequently-updated documents, e.g., lower level policies and procedures
- Points to the digital preservation plan for near-term priorities and timeframes
- Documents the policy approval and maintenance process

The framework includes one section for each of the seven attributes of a trusted digital repository: OAIS compliance, administrative responsibility, organisational viability, financial sustainability, technological and procedural accountability, system security, and procedural accountability. Some sections contain more than one component.

Framework Components

OAIS Compliance: consists of an explicit statement of the intent of the digital preservation program to comply with the Open Archival Information System (OAIS) Reference Model approved as ISO 14721 in 2003. The digital preservation plan delineates the specifics of OAIS compliance and the self-assessment results for the digital preservation program documents the status of the program's OAIS compliance. *Links:* strategic plan, preservation plan

Administrative Responsibility: makes an explicit commitment to digital preservation and to compliance with prevailing standards and practice.

> **Purpose:** makes explicit the intentions of an institution and defines the essential role a digital preservation program plays in fulfilling the mission to protect the organisation's digital assets. This section defines the rationale for the framework, identifies responsible parties and stakeholders, indicates the intended audience for the document, and places the document in the context of organiszation-wide efforts. The purpose statement might range from broad to narrow, reflecting the variations in intention for different types of digital archives. *Links:* mission statement, high-level policy statements, strategic plans.
> **Mandate:** stipulates the authority, jurisdiction, or governance upon which responsible parties have developed the digital preser-

vation program, e.g., laws, legislation, policies, and mission. This section may also address requirements that are not specifically identified as preservation, e.g., legal admissibility, authenticity, FOIA, ADA, Data Protection Acts, copyright legislation, public records acts, E-Government, National Grid for Learning (UK). *Links:* laws, legislation, contracts, policies, mission statement, regulations.
> **Objectives:** states the high-level aims and targets of the organisation for collecting, managing, preserving, and sustaining access to digital content. This section identifies the benefit of the program to an institution and its relationship to other objectives, goals, and policies. *Links:* strategic plans, goals and objectives, budgets, preservation plans, technology plans.

Organisational Viability: addresses the legal status as well as human and other resources needed to establish and maintain a digital preservation program.

> **Scope:** establishes the overall timeframe, levels of responsibility, boundaries, extent, limitations, and priorities of the digital preservation program. This section delineates what the organization's digital preservation program will do and, as importantly, will not do. The scope statement may be brief or extensive, depending on the nature of the program. The scope provides useful metric for measuring the effectiveness of the digital preservation program. *Links:* strategic plans, collection development policies, preservation plan, role definitions.
> **Operating Principles:** defines the key principles, models, processes, and assumptions upon which the digital preservation program is developed and implemented. This section is particularly important in establishing system-wide benchmarks for distributed programs when multiple operational and technical processes are implemented. Common principles include adherence to standards (in particular OAIS) and other accepted indicators of good practice, support for life cycle management, interoperability, evidence-based requirements, and preferred methods of preservation. *Links:* community and organisational good practice, workflow and process documents, procedures.
> **Roles and Responsibilities:** describes key stakeholders and their respective roles in digital preservation, including creators, producers, digital repository staff, administrators, financial managers, user groups, advisors, other repositories, and collaborators. This section makes an explicit statement that digital preservation is shared responsibility requiring participants within and beyond the organisation. It describes broad categories of roles and responsibilities and cites documents containing more specific descriptions.
> *Links:* role definitions with explicit responsibilities, documentation of current role assignments, job descriptions, organisational charts.
> **Selection and Acquisition:** provides the rationale and processes for developing and retaining collections based on specific parameters (e.g., formats, types of records, geographic scope). A clear articulation is critical to the success of a digital repository and ensures that collections support the institutional mission and priorities, and that requisite resources are made available for digital preservation. One aspect of auditing a digital archive is to verify that the stated mission and intended scope of a digital archive matches its actual content. Specific policies logically follow from

the conceptual statement in the framework to further collection development aspects, e.g., submission guidelines. *Links:* collection development policy, submission guidelines, ingest workflow.

Access and Use: identifies the designated communities for the digital preservation program and the barriers and/or restrictions to use of the digital content for which the digital preservation program is responsible. Specific policies should be developed to further articulate access and use requirements and restrictions. Note: A digital archive may be dark, dim, or lit, but the absolute proof of preservation is in the capability to provide meaningful long-term access. *Links:* access policy, deposit agreements, digital rights management rules and practice, user agreements.

Challenges and Risks: identifies the organisation's risks, difficulties, sense of urgency, and incentives for developing a digital preservation program. This section provides evidence that even though the full process may not be clearly understood, the need to act now is strong. *Links:* risk analyses, SWOT analyses, growth projections, examples of loss and near misses.

Financial Sustainability: documents the tangible basis for sustaining the digital preservation program.

Institutional Commitment: confirms and synthesizes the support for the program and the resources available to sustain the digital preservation program. *Links:* budgets, financial reports, fiscal policies, annual reports, succession plans, contracts

Cooperation and Collaboration: acknowledges that the organisation's effort exceeds or will exceed available resources and may not guarantee the safety of all vital assets. This section places the digital preservation programs into a broader context that recognizes the program's dependencies on other partners and on the community at large. Collaborations and partnerships may require formal, legally binding agreements that delineate explicit roles and responsibilities of each party. *Links:* partnership agreements, operating principles and practices for collaborative projects.

Technological and Procedural Suitability: this component summarizes the preservation approach, strategies and techniques that are employed by the digital preservation program to achieve stated objectives. This section states the general philosophy of the digital preservation program and points to relevant requirements, policies, standards, guidelines, and practice. It makes a tangible link to the preservation planning component of the digital preservation program and to the organisation's preservation plan. *Links:* preservation strategies, preservation plans for the organization and for specific content, deposit agreements.

System Security: specifies the organisation's commitment and approach to ensuring the accuracy, completeness, authenticity, integrity, and long-term protection of the organisation's digital assets. *Links:* security policies and procedures, disaster plan.

Procedural Accountability: acknowledges the need for and stipulates the means for ensuring the transparency and accountability of the digital preservation program's policies and operations.

Audit and Transparency: explicitly commits the organisation to periodic self-assessments and audits to evaluate, measure, and adjust the policies, procedures, preservation approaches, and practices of the digital preservation program. Transparency enables self-assessments and audits. Self-assessments and audits improve internal operations, facilitate external reviews, and contribute to the development of effective partnerships and collaborations. *Links:* audit and self-assessment schedules and results, strategic plans, preservation plans.

Framework Administration: describes the organisation's policies and practice pertaining to the development, approval, maintenance of the policy framework over time, e.g., frequency of updates and reviews, maintenance roles, expiration dates. The framework has little value if it has not received the appropriate approvals and has not been implemented. At minimum, the date and source of approval and the review cycle should be provided. *Links:* policy administration procedures, policy approval documentation.

Definitions: identifies terms and concepts that may be needed to understand the framework and may be instrumental in strategies for securing institutional commitment. This is an optional section, but one that can be very important. It is particularly important to include legally required and other mandated terminology and definitions. The section may either provide or point to requisite definitions. *Links:* definitions developed by the organisation, glossaries adopted by the organization.

References: provides citations for or pointers to key resources that were informed the development and application of the framework. This section identifies more detailed documents, both internal and external, that provide a deeper expression of the mission, underlying principles, illustrative processes, and sustaining roles. It may contain citations for these documents or point to a current list of relevant community standards and guidance. *Links:* cited resources, community lists of standards and practice.

Provenance of the Outline

This outline reflects the findings of the Cornell Digital Preservation Management workshop curriculum development project (co-developers, Anne R. Kenney and Nancy Y. McGovern, with funding from the National Endowment for the Humanities); lessons learned in the development of the Cornell University Library Digital Preservation Policy Framework (http://www.library.cornell.edu/iris/dpo/); and samples of policy frameworks developed by organisations that participated in the Cornell DPM workshop, e.g., the Library and Archives of Canada, N.C. State Library.

References

Open *Archival Information System (OAIS) Reference Model*, the January 2002 version is available at: http://nost.gsfc.nasa.gov/isoas/ref_model.html.

Attributes of a Trusted Digital Repository: Roles and Responsibilities, May 2002, available at: http://www.rlg.org/longterm/repositories.pdf.

Audit Checklist for Certifying Digital Repositories, January 2007 version available at: http://www.rlg.org/en/page.php?Page_ID=20769

Chapter 8

Project planning, management, quality assurance and evaluation

1. Introduction

The development of digital collections in South African libraries, archives and museums will take on an increasingly important role, especially as institutions realise the need to increase their visibility internationally. The Director or Deputy Director should maintain active control over this important work as it will involve the reallocation of resources, particularly human resources. Therefore a digital collections project should be developed and integrated into the operations of the library, archive or museum.

There is a direct correlation between the amount and quality of planning preceding and during a digital collections project and the success of the project. Planning helps to prevent inappropriate decisions from being taken. In addition, a well-planned project facilitates the management, quality assurance and evaluation of the project.

This chapter addresses the planning, management, quality assurance and evaluation of a digital collections project.

2. Project planning

The first step is to establish a small planning team whose primary responsibility is to plan and manage the implementation of the digital collection project. The team members should be chosen for their expertise in and knowledge of collection content, IT, intellectual property, and so on. The planning team will identify and evaluate alternatives and develop long-term goals.

The plan can also be used as a public relations or information document for senior management structures of the institution, governing boards and other interested individuals and groups and as a means of obtaining funds for the project.

The plan should be reviewed and revised periodically to reflect new ideas and changing conditions within the library and its environment. It should also include the library, archive or museum mission statement, general digital collections goals, and a long-term implementation strategy, as a minimum. If desired, a general statement for funding digital collections can also be included.

The planning process should result in the development of the mission and long-term goals of the digital collection.

2.1 The Long-term Goals for Digital Collections

There should be clear justification for developing digital collections in an institution, and the digitisation project must be compatible with the host organisation's wider mission in order to attract high level support from within the organisation. This should be stated in the library, archive or museum mission statement and goals.

Example: The mission of the library is to support teaching, learning and research in the university by providing access to the knowledge **generated** in the institution and through the provision of access to information and materials **not owned** by the institution, including archival collections.

The goals of the long-term digital collection could include the following:
- Improve access to the collection by making it available on the Internet and thus accessible to a wide range of communities.
- Provide access in innovative and exciting ways.
- Provide means to preserve original collections.
- Enhance searchability.
- Integrate different media.
- Organise, catalogue and index the collection.
- Help prevent and/or remedy disaster.
- Expose material which would otherwise have been 'hidden' or extremely difficult to access.
- Preserve and document the national collection through the use of digital surrogates, where appropriate.
- Enhance understanding and enjoyment of the visual arts through providing electronic access to digitised material, online and onsite.
- Provide improved access to unknown or little-used collections.
- Promote understanding of original works through improved indexing or some form of digital enhancement.
- Create resources that are tailored for use in learning and teaching.
- Ensure continued access to copies of fragile originals.
- Enhance the public knowledge, recognition or understanding of the collection.

2.2 Aims, Values and Outputs of the Project

The aim of a digital collection project should be identified. For example it might be to make the material available on the Internet and thus accessible to a wide range of communities and for preservation. Other aims could include organisation, cataloguing and indexing of the collection.

The outputs expected from the project should be stated. For example, the project aim may be to digitise a collection, and the final output would be for it to be available online and accessible for consultation *in situ*. Outputs associated with the project could include the preservation of the collection and the development of staff capacity and skills.

The planning process that will result in a plan requires various interactions that will be of benefit to the organisation.

2.3 Project Outline and Schedule

The phases of the project should be outlined along with the accompanying activities or tasks. The outline should be dynamic and could be altered as the project progresses and as new conditions and situations arise. The phases and tasks of a digitisation project include:

- Assessment and selection;
- The analysis and identification of the applicable processes and software, translation, re-keying, transcription and editing of OCR. These may be needed for some documents, depending on their condition;
- Copyright clearance or other research regarding rights and permissions. These should be obtained and rights and permissions metadata should be created;
- Preparation. This must include conservation assessment and/or treatment, if necessary.
- Design and implementation of document control;
- Standards definition and benchmarking for the different types of files to be created;
- Estimation of space requirements and confirmation of space availability;
- Digitisation. This includes scanning, OCR and creation of digital files (pdf, jpeg);
- Check-in and quality control for source materials and digital images; transfer of sources to original or new location; rehousing of materials, updating of catalogue records, as necessary;
- Creation of structural metadata;
- Creation of full text, including mark-up;
- File management: Loading of data to repository;
- Integration of digital images and metadata into an image database; hyperlinking associated catalogue records or other access points;
- Delivery. This can range from hand-crafting web pages to relying upon a highly automated system;
- Advertising, promotion, user evaluation;
- Long-term preservation of resources and digital files.

It is advisable to create a schedule for the completion of each of the tasks mentioned above, i.e.:
- An estimate of time (hours and days) required to complete the task;
- the resource person who will carry out the task.

Figure 8.1 Project Flow: Project flow diagram (Project outline)
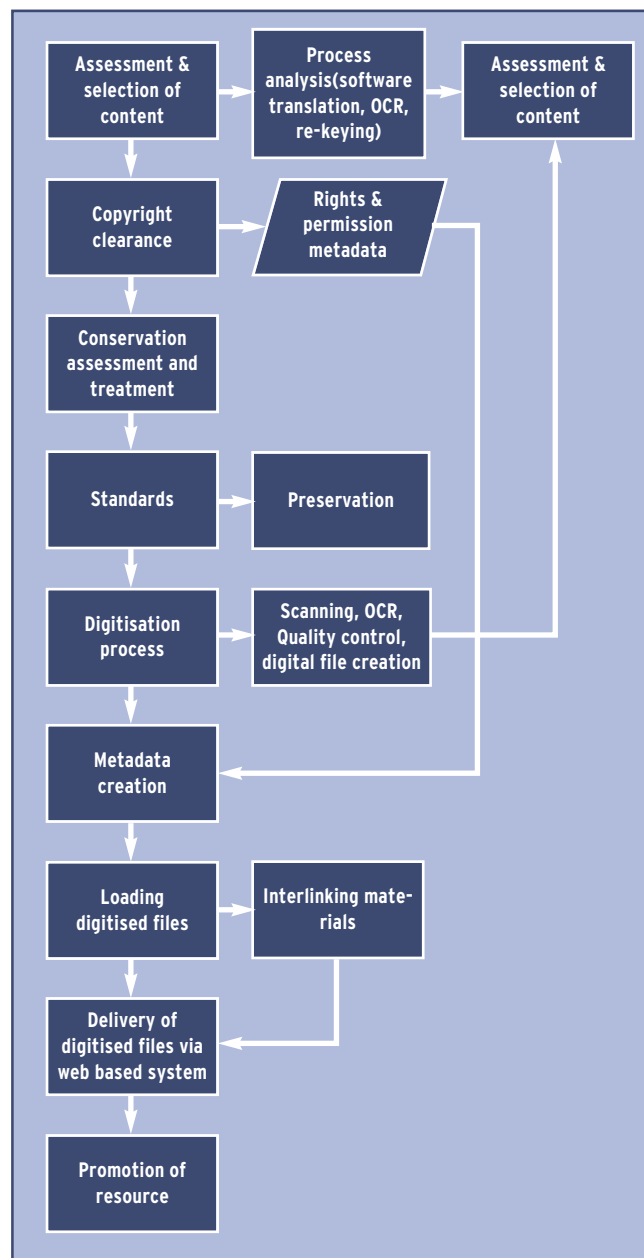
A software package could be used to create a project schedule. A Gantt chart or bar chart is a valuable tool to use for this activity. Spreadsheet packages can be used for document control and the measuring of progress.

2.4 Financial and Cost Planning

Developmental and operating costs are pertinent to the development of a digital collection.

Development costs are mostly once-off and include costs for the following:
- hardware,
- software,
- site preparation,
- network infrastructure,
- scanning of document (if outsourced),

- OCR of digital files,
- rekeying,
- staff training in Excel, Adobe and Scanning software,
- equipment – furniture, etc.

Operating costs are recurring costs and include:
salaries and wages of staff,
maintenance of hardware and software (refreshing and migration),
supplies.

In-house vs. Outsourcing:
It is important to consider whether materials should be digitised in-house or whether this function should be outsourced. Factors to consider include costs, condition of original material, and the skills set available in-house. Outsourcing may mean that skills, funds and equipment are not retained within the project, but may prove to be quicker, easier and more cost effective. If the decision is taken to outsource, the vendor must comply with the quality standards set for the project.

3. Project management

A project manager should be selected or appointed to lead the project in a project team approach or matrix approach. In the project team approach, staff needed to complete the project is assigned to the project manager for the duration of the project. In the matrix approach, staff remain in their functional units but contribute time to the digitisation project.

The project manager is responsible for the successful completion of the digitisation project. The manager ensures that milestone deadlines are met, and coordinates activities of all other participants. The project manager should carefully and regularly monitor the project, reviewing the achievement of the milestones and checking if all activities are completed on time, and that they are within the agreed quality standards and budget.

The project team should meet and discuss the progress of the project and any potential issues on a regular basis.

As deviations from the initial plan are identified, an analysis of the situation should be made and the decision on how to proceed should be taken by the project team. The project development plan should be updated accordingly and circulated to the implementing team and the relevant parties.

A pilot workflow should be carried out with half a dozen items to establish project cost for scanning, processing, metadata creation and quality control.

An effective project manager will gather and report production statistics, problem logs, feedback from staff, and expenditures.

Best practices for project management include:
- Adhering to the project schedule in a timely and cost-effective manner;
- Keeping a risk register;
- Keeping quality documentation throughout the project. This includes documenting the rationale, methodologies and activities,

systems staffing models, costs, as well as the lessons learned from a project;
- Wherever possible, building the delivery of some samples of intended outputs throughout the project rather than relying on delivery at the end. This can help to engage potential audiences at an early stage. This also helps to test the processes involved and deal with quality issues with suppliers at an early stage;
- Ensuring that people who work on the project have a meaningful experience.

4. Quality assurance

The principle of quality means maintaining and applying digitisation standards, both in the sense of specific expectations and requirements that should be complied with, and the ideals of excellence that should be aimed at. Applying the principle of quality entails evaluating services and products against set standards, with a view to improvement, renewal or progress.

The standards should be established during the planning stage and implemented throughout the project. Every project will need to set quality standards that reflect the aims and objectives of the project. A designated quality officer and not the creator of the work should sign off completed tasks. Any error or faults should be rectified with minimal delay.

A Quality Assurance (QA) Project Plan should cover the entire project from planning, through implementation to assessment. The following template of area and indicators can be used in the QA process:

Figure 8.2 Quality Assurance Outline

AREA	INDICATORS	
	Standards	
	Best practices	
	Evidence of compliance	

5. Evaluation

The purpose of evaluation is to measure and report the success of the digitisation programme in meeting objectives, and to inform decisions in designing future programmes and projects. The evaluation should be part of the project plan from the beginning. Consult JISC Project Management Guidelines for more information.

6. Communication plan/progress report

A regular progress report should be produced and made available to various stakeholders. This would assist is raising awareness about the programme objectives.

7. Conclusion

The relocation of resources that is demanded by the introduction and maintenance of a digital collection makes careful planning and management vitally important. Not only is the planning for the initial establishment of a digital collection important, but the long-term goals must also be mapped out. However the plan must not be inflexible and

should allow for unforeseen developments. This is why it is essential for an institution to establish a planning team that can guide the process and keep careful records of what has taken place. Changes in the planning team must also be expected, and contingency arrangements should be part of the plan. Figure 8.1 provides an idea of the different stages of the project and illustrates the complexity of such a project. The chapter further illustrates the various aspects that must be taken into account. Continuous quality assurance, evaluation and reporting will ensure the success of the project.

References

Distributed National Electronic Resource (DNER). Working with the Distributed National Electronic Resource (DNER): Standards and Guidelines to Build a National Resource. (2001) Available from: http://www.jisc.ac.uk/uploaded_documents/ACF127.pdf (accessed 15 February 2009)

JISC Digital Media – Crossmedia. Managing a Project website. Available from: http://www.jiscdigitalmedia.ac.uk/crossmedia/docs/category/managing-a-project/ (accessed 15 February 2009)

JISC. *Evaluation of the JISC Digitisation Programme Phase 1 and International Contextualisation*. (nd) Available from: http://www.jisc.ac.uk/whatwedo/programmes/digitisation/evaluation/lessons.aspx (accessed 20 March 2009)

JISC. *JISC Project Management Guidelines*. (2008) Available from: http://www.jisc.ac.uk/media/documents/funding/project_management/projectmanagementguidelines.pdf (accessed 22 March 2009)

JISC. *Project planning* website. Available from: http://www.jisc.ac.uk/fundingopportunities/projectmanagement/planning.aspx (accessed 8 March 2009)

National Gallery of Australia. (2006). *Collections Digitisation Strategy*. Available from: http://nga.gov.au/Collection/forms/Collections DigitisationStrategy.pdf (accessed 22 March, 2009)

National Information Standards Organisation. *A Framework of Guidance for Building Good Digital Collections*, 3rd edition. (2007). Available from: http://www.niso.org/publications/rp/framework3.pdf (accessed 2 March 2009)

Northeast Document Conservation Center. *Handbook for Digital Projects: A Management Tool for Preservation and Access*. (2000) Available from: http://www.nedcc.org/resources/digitalhandbook/dman.pdf (accessed 8 March 2009)

Mind Tools Ltd. *Gantt Charts*. (nd) Available from: http://www.mindtools.com/pages/article/newPPM_03.htm (accessed 20 March 2009)

UKOLN. *Implementing Your Own QA website*. Available from: http://www.ukoln.ac.uk/qa-focus/documents/briefings/briefing-58/ (accessed 12 March 2009)

UKOLN. *The QA Focus* website. Available from: http://www.ukoln.ac.uk/qa-focus/ (accessed 15 February 2009) University of Hertfordshire. Heds Digitisation Services. *Digitisation: Strategic and Management Issues*. (2009) Available from: http://heds.herts.ac.uk/resources/Papers/HEDSITForum.pdf (accessed 12 March 2009)

University of Central England. Evidence Base. *Evaluation of the JISC Digitisation Programme Phase 1 and International Contextualisation*. Available from: www.jisc.ac.uk/media/documents/programmes/digitisation/phaseoneevaluation.doc (accessed 15 February 2009)

Appendix A

Glossary

Archival Information Collection (AIC)
An Archival Information Package whose Content Information is an aggregation of other Archival Information Packages. Source: OAIS

Archival Information Package (AIP)
An Information Package, consisting of the Content Information and the associated Preservation Description Information (PDI), which is preserved within an OAIS. Source: OAIS

Archival Storage
The OAIS entity that contains the services and functions used for the storage and retrieval of Archival Information Packages. Source: OAIS

ASCII
Type of coding by computers often used in diacritical symbols

Authenticity
Assurance that Provenance and appearance are not tampered with

Bandwidth
The amount of data that can be carried from one point to another in a given time period

Binary
Switches being either in a "1" and "on" position/value or in a "0" and "off' position/value

Bit
One binary digit (a 1 or a 0)

Byte
A sequence of 8 bits

Born Digital
A resource that is created in an electronic format. A document created on a word-processor is an example of a born digital resource.

Business Continuity
Describes the processes and procedures an organisation puts in place to ensure that essential functions can continue during and after a disaster. [1] A note regarding preservation: "Backups vs Preservation: Disaster recovery strategies and backup systems are not sufficient to ensure survival and access to authentic digital resources over time. A backup is a short-term data recovery solution following loss or corruption and is fundamentally different to an electronic preservation archive.[2] Source: There are many definitions for business continuity. [1] SearchStorage.com. [2] Continued access to authentic digital assets, JISC Digital Preservation Paper, Nov 26, 2006

Calibrate
An action to adjust instrumentation (e.g. computer) precisely to perform a particular function

Codec
Compression/decompression of large downloadable files

Computer protocol
Standardised method of information transfer between electronic devices

Controlled Vocabulary
A limited set of accepted terms used to describe a resource

Digital Collection
A related group of digital items

Data Curation
Value-added activities and features that stewards of digital content engage in to make digital content meaningful or useful; this term may specifically refer to specific types of research data or to digital content more generally

Data object
A digital object (can also be a physical object)

Digital curation
The process of establishing, maintaining and developing long term repositories of digital assets for ongoing access

Digital Preservation
Activities that are undertaken by a digital curator to ensure that the digital content for which the digital curator has responsibility is maintained in usable formats over time and can be made available in meaningful ways to current and future users. Source: ICPSR

Digital Resource
Refers to the digital derivative of an existing physical object or a born digital resource. A digital image is an example of a digital resource

Digital Stewardship
Contributions to the longevity and usefulness of digital content by its caretakers at any point in the lifecycle of digital content management over time

Digitisation
Conversion of analogue (physical material) by a scanner or other electronic device to a machine readable format

Dissemination Information Package (DIP)
The Information Package, derived from one or more AIPs, received by the user in response to a request to the repository

Emulation
A strategy for continuing access to digital materials that mimics or re-creates the digital object's original technical environment using current technology

Flash
File format to deliver moving images on electronic devices

Formats
The software created to "carry" specific data types (bitstreams)

Interoperable
Able to operate in conjunction and accessible across different platforms and hardware

JPEG2000
Wavelet-based image compression standard

Lossless
A class of data compression algorithms that allows for the exact original data to be reconstructed from the compressed data

Lossy format
Unnecessary data is disregarded by the compression software

Master image
Originally created document or surrogate of physical

Metadata
A term that refers to structured data about data. Metadata is an old concept (e.g., card catalogs and indexes), but metadata is often essential for digital content to be useful and meaningful. Metadata can capture general or specific information about digital content that may define administrative, technical, or structural characteristics of the digital content. *Preservation metadata* is a specific sub-set of metadata that documents the lifecycle of digital content from creation through processing, storage, preservation, and use over time. Preservation metadata is required at the aggregate (e.g. collection level) and at the item (e.g. file level). All preservation actions that are applied to digital content over time should be captured in preservation metadata, for example. Source: ICPSR

Migration
Upgrading or transfer of document to a later document version or hardware equipment

OAIS
The Open Archive Information System (OAIS) Reference Model, an ISO standard that formally expresses the roles (producer, management, consumer, and implicitly archives), functions (common services, ingest, archival storage, data management, administration, preservation planning, and access), and content (submission information package, archival information collection, archival information package, and dissemination information package) of an archive. It was approved as an ISO standard in 2003, and was updated in May 2009. Source: OAIS

Opt-out
Method to state that information will be removed if copyright does imply

Raster image
Digital image composed of pixels

Repository
An organisation that maintains information for access and use.

Search engine
A tool designed to search for information on the World Wide Web

Submission Information Package (SIP)
An Information Package that is delivered by the producer to the repository for use in the construction of one of more AIPs.

Vector image
Digital image composed of paths

Appendix B

Acronyms

AACR	Anglo American Cataloguing Rules
AAT	Art and Architecture Thesaurus
AIC	Archival Information Collection
AIP	Archival Information Package
ARC-IA	Internet Archive Format ARC File Format
AVI	Audio Video Interleave (multimedia container format)
CAD	Computer-aided design
CAM	Computer-aided manufacturing
CD	Compact disc
CCSDS	Consultative Committee for Space Data Systems
CLIR	Council on Library and Information Resources
CNRI	Corporation for National Research Initiatives
CRL	Center for Research Libraries
CRT	Cathode Ray Tube
DC	Dublin Core
DCC	Digital Curation Centre, United Kingdom
DCMES	Dublin Core Metadata Element Set
DIP	Dissemination Information Package
DOI	Digital Object Identifier
DRAMBORA	Digital Repository Audit Method Based on Risk Assessment Toolkit
DVD	Digital Video Disk
EAD	Encoded Archival Description
FRAD	Functional Requirements for Authority Data
FRBR	Functional Requirements for Bibliographic Records
GIF	Graphic Interchange Format (used in vector based images)
GIS	Geographic Information System
HTML	Hypertext Markup Language
JPG/JPEG	Joint Photographic Experts Group (format for image compression)
JPEG2000	Joint Photographic Experts Group 2000 format
IAM	Identity and Access Management
ICPSR	Inter-university Consortium for Political and Social Research
IEC	International Electrotechnical Commission
iPres	International Conference on the Preservation of Digital Objects
IS&T	Society for Imaging Science and Technology
ISO	International Standards Organisation
JPEG	Joint Photographic Experts Group (format for image compression)
LAMP	Linux / Apache / MySQL / PHP,Perl,Python
LCSH	Library of Congress Subject Headings
LDAP	Lightweight Directory Access Protocol
LCD	Liquid Crystal Display
LIFE	Life Cycle Information for E-Literature Project, United Kingdom
MARC	Machine-Readable Cataloguing Format
MeSH	Medical Subject Headings
METS	Metadata Encoding and Transmission Standard
MIME	Multipurpose Internet Mail Extensions
MIX	Technical Metadata for Digital Still Images Standard
MODS	Metadata Object Description
MP3	MPEG-1 Audio Layer 3 (lossy digital audio encoding format)
MP4/MPEG-4	Moving Picture Expert Group-4 (storage format for moving pictures)
NDIIPP	National Digital Information Infrastructure and Preservation Program
NISO	National Information Standards Organisation (USA based)
OAI	Open Archives Initiative
OAI-PMH	Open Archives Initiative Protocol for Metadata Harvesting
OAIS	Open Archival Information System
OCR	Optical Character Recognition
OS	Operating System
PADI	Preserving Access to Digital Information
PDF	Portable Document Format
PNG	Portable Network Graphics (widely used in web pages)
PLANETS	Preservation and Long-term Access through Networked Services
PNG	Portable Network Graphics
PREMIS	Preservation Metadata Implementation Strategies
PURL	Persistent Uniform Resource Locators
RAD	Rules for Archival Description
RDA	Resource Description and Access
RDBMS	Relational Database Management System
RLG	Research Libraries Group
RSS	Rich Site Summary (delivering regularly changing web content)
RTF	Rich Text Format
SAN	Storage Area Network
SCORM	Sharable Content Object Reference Model
SGML	Standard Generalized Markup Language
SIP	Submission Information Package
SRB	Storage Resource Broker
SUM	Service Usage Model
SVG	Scalable Vector Graphics
SWORD	Simple Web Service Offering Repository Deposit
TDR	Trusted Digital Repository
TEI	Text Encoding Initiative
TGN	Thesaurus of Geographic Names
TIFF	Tagged Image File Format (mostly used for master images)
TRAC	Trustworthy Repositories Audit and Certification
URI	Uniform Resource Indicator
URN	Uniform Resource Names
URL	Uniform Resource Locator
UUID	Universally Unique Identifier
WARC	Web Archive File Format
WAV	Waveform audio format (used for storing audio bitstream)
WIPO	World Intellectual Property Organization
XENA	XML Electronic Normalising for Archives
XML	eXtensible Markup Language

Appendix C

Useful resources

- African Copyright & Access to Knowledge Project (ACA2K) – Copyright & A2K Issues Blog http://kim.wits.ac.za
- African Digital Library http://www.africandl.org.za/about.htm
- African Journals Online http://www.ajol.info
- African Online Digital Library http://www.aodl.org
- Ariadne (UK) http://www.ariadne.ac.uk
- Copyright Clearance and Digitisation in UK Higher Education: Supporting Study for the JISC/PA Clearance Mechanisms Working Party http://www.ukoln.ac.uk/services/elib/papers/pa/clearance/ study.doc
- CSIR's Research Space http://researchspace.csir.co.za/dspace/
- Database of African Theses and Dissertations (DATAD) http://www.aau.org/
- Digital Innovation of South Africa (DISA) http://www.disa.ukzn.ac.za
- Digital Preservation http://www.lockss.org
- Digital Resource Management Workshop hosted by DISA http://www.disa.ukzn.ac.za/index.php?option=com_content&view=article&id=122&Itemid=107
- Digital Rights Management and Access to Information: a developing country's perspective http://libres.curtin.edu.au/libres19n1/index.htm
- Directory of Access Repositories http://www.opendoar.org/
- Directory of Open Access Journals http://www.doaj.org
- DISA Dublin Core Metadata Generator http://www.disa.ukzn.ac.za/index.php?option=com_wrapper&view=wrapper&Itemid=89
- DISA Guidelines for Best Practice http://www.disa.ukzn.ac.za/index.php?option=com_docman&task=cat_view&gid=62&Itemid=88
- Dspace http://www.dspace.org
- DSpace Technical Workshop, 7 – 11 September 2009, hosted at Stellenbosch University: http://www.lib.sun.ac.za/dspace/workshop and http://ir.sun.ac.za/wiki
- eIFL Handbook on Copyright and Related Issues for Libraries http://www.eifl.net/cps/sections/services/eifl-ip/issues/handbook/handbook-complete-text
- Electronic Information for Libraries (eIFL) www.eifl.net/
- Eprints http://www.eprints.org/software/
- Fair Dealing in an Electronic Environment (UK) http://www.ukoln.ac.uk/services/elib/papers/pa/fair/intro-old.html#guidelines http://www.aca2k.org
- Fedora http://www.fedora-commons.org/
- Framework of Guidance for Building Good Digital Collections http://framework.niso.org/
- Good practice http://www.diglib.org/standards/bmarkfin.htm
- Hillman, Diane I. Getting the word out: making digital project metadata available to aggregators. First Monday, Vol 11, No 8, 2006. http://www.uic.edu/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/1385/1303
- How to install DSpace 1.5.2 on Linux using Ubuntu 8.04LTS (online tutorial) http://ir.sun.ac.za/wiki
- IFLA Committee on Copyright and Other Legal Matters http://www.ifla.org/III/clm/copyr.htm
- IFLA Information Policy: Copyright and Intellectual Property http://www.ifla.org.sg/II/cpyright.htm
- IFLA Position Paper on Copyright in the Electronic Environment http://www.ifla.org/V/press/pr961002.htm
- IFLA/IPA - Publishers and librarians promote common principles on copyright in the electronic environment http://www.ifla.org/V/press/ifla-ipa.htm
- ILFA- Division of Bibliographic Control http://www.ifla.org/VII/d4/dbc.htm
- Inquiring Librarian Blog http://inquiringlibrarian.blogspot.com/
- International Network for the availability of Scientific Publication (INASP) http://www.inasp.info
- International Study on the Impact of Copyright Law on Digital Preservation www.digitalpreservation.gov/library/resources/pubs/docs/
- Introducing Copyright: a plain language guide to copyright in the 21st century (Commonwealth of Learning) http://www.col.org/resources/publications/monographs/Pages/Copyright.aspx
- Library of Congress American Memory http://memory.loc.gov/ammem/about/index.html
- OAI for beginners: the Open Archives Forum online tutorial http://www.oaforum.org/tutorial/english/intro.htm
- Open Access related workshops and conferences http://www.sivulile.org/workshops
- Open Review of South African Copyright Act and Copyright Resources http://www.shuttleworthfoundation.org/node/599 http://copyright.shuttleworthfoundation.org/wiki/Resources
- Partnering on Copyright (JISC/SURF) http://info.lut.ac.uk/departments/ls/disresearch/poc/pages/otherresources-programme.html

- PASA - Publishers' Association of South Africa FAQ's http://www.publishsa.co.za/home.php?cmd=copy_faq
- Ranking Web of World Repositories http://repositories.webometrics.info/about_rank.html
- Selecting for Digitization and Copyright Issues http://www.cdlib.org/inside/diglib/resources/onres_dig_selection.html#copyright
- SHERPA/RoMEO: Institutional Repositories www.sherpa.ac.uk/repositories
- SHERPA/RoMEO: Publishers allowing the deposition of their published version/PDF in Institutional Repositories http://www.sherpa.ac.uk/romeo/PDFandIR.html
- SHERPA/RoMEO: Publishers' copyright policies & self-archiving http://www.sherpa.ac.uk/projects/sherparomeo.html
- SHERPA-JULIET: Research funders' open access policies http://www.sherpa.ac.uk/juliet
- Shreeves, Sarah L. Moving towards shareable metadata. First Monday, Vol 11, No 8, 2006. http://www.uic.edu/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/1386/1304
- South Africa : Copyright Act No. 98 of 1978 (as amended in 2002) http://portal.unesco.org/culture/en/files/30224/11416532713dz_copyright_2002_en.pdf/dz_copyright_2002_en.pdf
- South Africa. Copyright Regulations, 1978 (Section 13) http://www.wipo.int/clea/en/text_html.jsp?lang=EN&id=4069
- South Africa: Copyright Act (Consolidation), 20/06/1978 (1992), No. 98 (No. 125) http://www.wipo.int/clea/en/details.jsp?id=4067
- South Africa: Copyright (Cinematograph Films), Regulations, 24/10/1980, No. R2140 http://www.wipo.int/clea/en/details.jsp?id=4071
- South Africa: Intellectual Property, Amendment Act, 1997 http://www.wipo.int/clea/en/details.jsp?id=4092
- South African National Library and Information Consortium (SANLIC) http://www.cosalc.ac.za
- Summary of OAI Metadata Best Practices http://www.diglib.org/architectures/oai/imls2004/training/metadataFinal.pdf
- SURF Foundation Copyright Toolkit http://copyright.surf.nl/copyright/implementing_principles/agreements_publisher_author/copyright_toolbox.php
- SURF Foundation University Copyright Policies http://copyright.surf.nl/copyright/implementing_principles/university_copyright_policies/
- Trusted Digital Repository http://www.oclc.org/programs/ourwork/past/trustedrep/repositories.pdf
- Trusted Digital Repositories http://www.icpsr.umich.edu/dpm/dpm-eng/eng_index.html
- University of the Witwatersrand, Johannesburg. Library Copyright Information http://web.wits.ac.za/Library/Services/COPYRIGHT.htm
- University of the Witwatersrand, Johannesburg. Library Copyright Portal http://web.wits.ac.za/Library/ResearchResources/SubjectPortals/Copyright+and+Related+Issues.htm
- WIPO Study on Copyright Limitations and Exceptions for Libraries and Archives http://www.wipo.int/meetings/en/doc_details.jsp?doc_id=109192
- WIPO Study on Limitations and Exceptions in the Digital Environment http://www.wipo.int/meetings/en/doc_details.jsp?doc_id=16805
- Working with the Distributed Electronic Resource (DNER): Standards and Guidelines to Build a National Resource http://www.jisc.ac.uk/uploaded_documents/ACF127.pdf
- Workflow http://projects.exeter.ac.uk/charter/documents/DigitisationWorkflowGuidev5.pdf
- World Wide Science www.worldwidescience.org